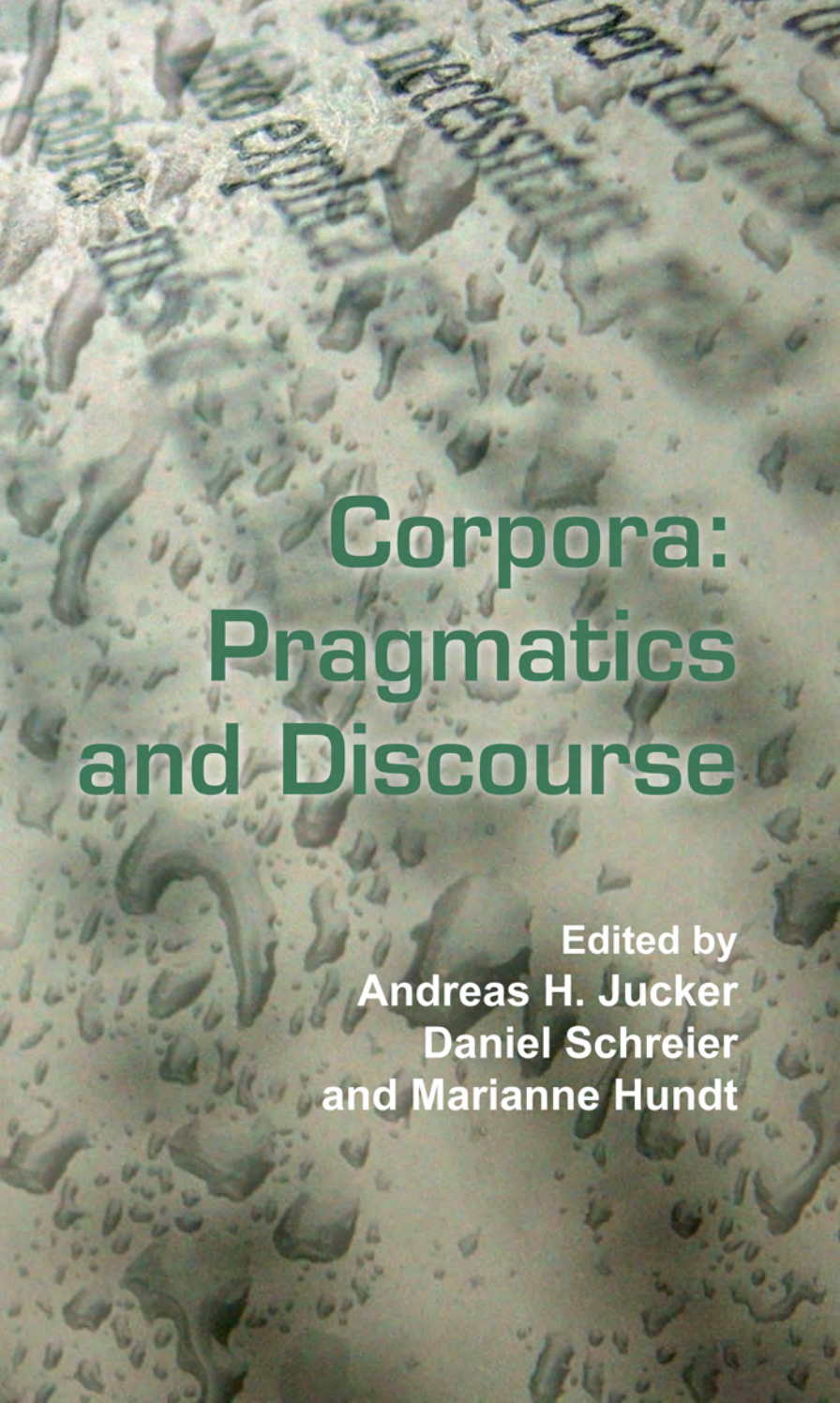
Corpora:
Pragmatics
and Discourse
Edited by
Andreas H. Jucker
Daniel Schreier
and Marianne Hundt

Corpora: Pragmatics and Discourse

LANGUAGE AND COMPUTERS: STUDIES IN PRACTICAL LINGUISTICS

No 68

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

Corpora: Pragmatics and Discourse

Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)

Ascona, Switzerland, 14-18 May 2008

Edited by
Andreas H. Jucker
Daniel Schreier
and Marianne Hundt

Rodopi

Amsterdam - New York, NY 2009

Cover design: Pier Post

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-2592-9

Printed in The Netherlands

Contents

Introduction

Corpus linguistics, pragmatics and discourse

Andreas H. Jucker, Daniel Schreier and Marianne Hundt

University of Zurich

The continuing success story of corpus-based linguistic research manifests itself in various guises. Certainly the most visible of these is the ever-growing number of monographs, edited volumes and research articles in peer-reviewed international journals, as witnessed by the continued publication of textbooks (Teubert and Čermáková 2004; McEnery, Xiao and Tono 2006; O'Keeffe, McCarthy and Carter 2007; Mukherjee 2009) and the recent publication of a major handbook (Lüdeling and Kytö 2008). The proliferation of available corpora (see McEnery, Xiao and Tono 2006: 59-70; and Mukherjee 2009: 41-64 for relevant overviews) attests to the ongoing innovation in the field, and they are another important cornerstone since they provide new databases for the research community. Last but not least, a central factor in the advancement of corpus linguistics as a discipline in its own right is the annual platform available in the form of international conferences. For English corpus-based linguistics, the most important of these is the one organised by members of the *International Computer Archive of Modern and Medieval English* (ICAME). The 29th conference was held in Ascona (Switzerland) in May 2008. Each conference has its own theme, and the one chosen for the 2008 conference was *Corpora: Pragmatics and Discourse*. The 22 papers in this volume, selected from a much larger number submitted for publication, go back to presentations given at this conference.

The success story of corpus linguistics can also be explained by the fact that its tools are being applied to an ever-growing range of linguistic research questions, so that corpus linguistic research methods have become mainstream in many, perhaps even most, subfields of linguistics: corpora are increasingly used even by formal linguists (cf. some of the papers in Kepser and Reis 2005; or Bresnan et al. 2007a) or psycholinguists (cf. Baayen 2007; for a combination of a formalist syntactic approach and psycholinguistic aspects with corpus linguistic methodology, see Bresnan et al. 2007b). At the other end of the spectrum, we find qualitative pragmatic research (see, for instance, the special issue of *Journal of Pragmatics*, Mey 2004, devoted to corpus linguistics; or the recent collection of papers in Romero-Trillo 2008a). In choosing *Corpora: Pragmatics and Discourse* as the conference theme, ICAME 29 emphasised the interdisciplinary character of this approach to linguistics and its potential for other fields that await exploration. The goal was to show possible future applications of corpus methods and to thus encourage increased involvement of corpus linguists in a research field that for a long time has been peripheral.

In a very general sense, pragmatics is about how language is used in actual situations. It is concerned with the ways in which speakers and hearers cooperate to negotiate meanings. In a broader conceptualization often adopted by European

researchers, pragmatics also covers the social and cultural conditions of language use (see, for instance, Verschueren 1999), while in the more restricted cognitive-inferential conceptualization of pragmatics, which is more common among Anglo-American researchers, pragmatics focuses on implicit and non-literal meanings that are derived from the use of language (see, for instance, Cruse 2000). However, both strands of pragmatics, in spite of their empirical orientation, have generally avoided large-scale corpus-based investigations for a number of reasons. For empirically-minded pragmaticists, the focus has traditionally been on individual conversations and other types of data amenable to manual analysis. Pragmatics is interested in how language is used in actual situations and focuses on the functions of specific linguistic elements. But speech functions do not lend themselves to easy searches in large computerized corpora for methodological reasons, since they have to be identified in a one-by-one fashion by the analyst.

So far, only a small number of pragmaticists have used large corpora to retrieve their data. Karin Aijmer, for instance, used the *London-Lund Corpus of Spoken English* to describe conversational routines, such as thanking, apologizing and making requests (Aijmer 1996) or discourse particles like *now*, *oh*, *ah*, *just* and *sort of* (Aijmer 2002). Similarly, Deutschmann (2003) used the *British National Corpus* to investigate apologizing in British English. Stenström and her research associates compiled their own corpus, the Bergen *Corpus of London Teenage Language* (COLT) for pragmatic analyses of the language of teenagers (see Stenström, Andersen and Hasund 2002). Andersen (2001) investigated how London teenagers use pragmatic markers to express their stance or propositional attitudes.

These research areas pose considerable methodological challenges. In all of these studies, the starting point is either a discourse particle with a fixed form that can easily be retrieved from a large corpus, or a speech function that is generally realized in a small number of variant patterns. Andersen searched for the discourse marker *like* and tags such as *innit* to investigate their function in expressing propositional attitudes. Aijmer searched for typical items for thanking, apologizing and requesting, such as *thank you* and *thanks*; *sorry, pardon* and *excuse*; and *I want you to* or *could you*, to mention just a few relevant examples. Deutschmann (2003) argues that apologies in British English tend to be realized in just a few routinized forms containing standard elements, such as *sorry*, *excuse* or *pardon*, and while it is impossible to search directly for the speech function of an apology, it is a straightforward task to search for such standard elements. In the meantime, these pioneering ideas have been extended to other speech acts, such as directives (see Vine 2004 for directives in contemporary New Zealand English; Kohnen e.g. 2000, 2004, 2008 for directives in Old and Middle English), requests (Wichmann 2004), promises (Valkonen 2008) and compliments (Jucker et al. 2008).

The term 'discourse' creates a somewhat different research focus. In a very general sense, it can be understood to refer to "language above the sentence or above the clause" (Stubbs 1983: 1). In this conceptualization, one might talk about the discourse structure of classroom interaction or the discourse structure of

news interviews with specific sequences of elements. In another sense, the term discourse can be used to refer to the totality of linguistic practices that pertain to a particular domain or that create a particular object (cf. Baker 2006: 4). In this sense we can talk of the discourse of football or the "discourse of compulsory heterosexuality" (Sunderland 2004: 55; Baker 2006: 5). Baker adopts this wider use of the term discourse in his introduction to corpus-linguistic methods in discourse analysis.

Both approaches to discourse have been combined with corpus-linguistic methods in recent years. Botley and McEnery (2000), for instance, adopt a narrow view in their volume on corpus-based and computational approaches to discourse anaphora. Partington et al. (2004) include papers both on the local organization of specific types of discourse (e.g. Biber et al. 2004 on discourse units in university registers) and on discourse in the wider sense (e.g. the paper by Garzone and Santulli 2004 on the potential of corpus linguistics for critical discourse analysis). Sanderson (2008), to mention one more example, investigates academic discourse in English and German research articles and examines discourse construction in a very wide sense, including social interaction, identity construction and metadiscourse.

The point we would like to make here, and in fact the essence of the reason why *Corpora: Pragmatics and Discourse* was chosen as the theme of ICAME 29, is that the potential of corpus linguistics has not yet been fully explored for either discourse analysis or pragmatics. Though both disciplines involve research on language usage, very few corpus linguists have tackled research questions in the fields of pragmatics and discourse, and only a handful of pragmaticists (see above) have applied corpora as a tool of analysis. We fully agree with Partington et al.'s (2004) assessment that:

> For some considerable time, then, the dichotomy was virtually complete: corpus linguists were generally unaware that their quantitative techniques could have much to say about discourse, while discourse analysts rarely saw reason to venture forth very far from their qualitative ivory tower. However, over the last decade, a number of developments took place, both technical and philosophical, which gradually made it possible to contemplate the mating of discourse and Corpus Linguistics. (Partington et al. 2004: 13)

We hope that ICAME 29 and this conference volume will help redress this sense of dichotomy and bridge the gap between corpus linguistics, pragmatics and discourse analysis. This may seem ambitious, but it is striking that other well-established fields of linguistics have, over the last 20 years, used corpus-linguistic methods extensively, to the extent that a growing number of linguists who do not consider themselves primarily as corpus linguists now have training and background knowledge in the compilation and analysis of corpora. This is perhaps most evident in the case of diachronic linguistics, which has received a major boost by the opening-up of new avenues of research via the application of

corpus-linguistic tools (cf. the *Helsinki Corpus*; ARCHER, *A Representative Corpus of Historical English Registers*, etc.) or in the case of variationist sociolinguistics, where large corpora are compiled with the aim of analysing language variation and change (to name but one, the *Toronto English Corpus*, compiled by Sali Tagliamonte and her associates at the University of Toronto, which has a size of 1.8 million words from a total of 214 speakers, male and female and aged between 8-92).

Given such fruitful expansion of corpus linguistics to other fields, the time is right to encourage and promote more systematic cooperation between researchers investigating pragmatics and discourse on the one hand and those working with corpus-linguistic methods on the other. Romero-Trillo (2004b) calls the co-operation between corpus linguistics and pragmatics a "mutualistic entente" with the intentional allusion to partners who form alliances in bellicose situations in order to have a stronger position against external forces: "pragmatics and corpus linguistics have not only helped each other in a relationship of mutualism, but, they have also made common cause against the voices that have derided and underestimated the utility of working with real data to elucidate the patterns of language use" (Romero-Trillo 2008b: 1).

The main aim of this volume is to show that both sides profit from such an interaction and to outline areas where this seems particularly promising. Pragmatics and discourse analysis will both benefit from establishing and improving rigorously empirical research methods that offer further insights into old and new research questions, which are then tested and refined against large sets of data of every-day usage. Corpus-linguists, in turn, are also likely to profit from a different range of research questions which provide new challenges.

Work in the early years of corpus linguistics tended to concentrate on syntactic, morphological and lexico-grammatical patterns and stylistic issues that were amenable to transformations into specific search strings. Of these, stylistic studies which cast their focus beyond the sentence (cf. Biber 1988) could even be seen as an early bridge between traditional pragmatics and corpus linguistics. In the meantime, the corpora available have mushroomed. The early all-purpose corpora have given way to a wealth of single genre corpora, such as the *Corpus of Early Modern English Tracts*, the *Corpus of Early English Correspondence*, the *Zurich English Newspaper* corpus or the *Corpus of English Dialogues*, to mention a few with a historical orientation. In addition, linguists are exploring the usefulness of text databases (i.e. electronically available newspapers or collections of fictional writing), the citation base of the *Oxford English Dictionary* and the rich resources on the internet for corpus linguistic studies. And the research questions that are tackled with corpus linguistic methods extend to fields of linguistics which hitherto shied away from using large computerized corpora for their investigations, including even the investigation of thought styles.

A considerable number of papers presented at the 29th ICAME conference in Ascona (Switzerland) focussed on pragmatic and discourse analytic issues in response to the theme of the conference. This is reflected in this volume, which, therefore, bears the conference topic as its title.

The first section of the volume comprises papers on the special topic of the conference: pragmatics and discourse. It opens with two plenary papers, one by Thomas Kohnen and one by Irma Taavitsainen. Both of them apply a corpus pragmatic perspective to historical data. The following papers continue the historical corpus pragmatic perspective, while the remaining papers of the first section are devoted to the pragmatics and discourse of present-day varieties of English.

The second part of the volume extends the perspective to case studies on specific problems of the English lexicon, grammatical aspects and semantics. The last section, finally, comprises three papers that are, in a sense, more technical. While many papers in this volume deal to some extent with theoretical issues of corpus compilation and analysis, the three papers in this last section focus specifically on such methodological problems.

References

Aijmer, K. (1996), *Conversational Routines in English. Convention and Creativity*. London: Longman.

Aijmer, K. (2002), *English Discourse Particles. Evidence from a Corpus*. Studies in Corpus Linguistics 10. Amsterdam and Philadelphia: John Benjamins.

Andersen, G. (2001), *Pragmatic Markers and Sociolinguistic Variation. A Relevance-theoretic Approach to the Language of Adolescents*. Pragmatics & Beyond New Series 84. Amsterdam and Philadelphia: John Benjamins.

Baayen, R.H. (2007), 'Storage and computation in the mental lexicon', in: G. Jarema and G. Libben (eds.) *The Mental Lexicon: Core Perspectives*. Oxford and Amsterdam: Elsevier. 81-104.

Baker, P. (2006), *Using Corpora in Discourse Analysis*. London: Continuum.

Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., E. Csomay, J.K. Jones and C. Keck (2004), 'Vocabulary-based discourse units in university registers', in: A. Partington, J. Morley and L. Haarman (eds.) *Corpora and Discourse*. Studies in Language and Communication 9. Bern: Peter Lang. 23-40.

Botley, S. and A.M. McEnery (eds.) (2000), *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam and Philadelphia: John Benjamins.

Bresnan, J., A. Deo and D. Sharma (2007a), 'Typology in variation: a probabilistic approach to *be* and *n't* in the survey of English dialects', *English Language and Linguistics*, 11(2): 301-346.

Bresnan, J., A. Cueni, T. Nikitina and H. Baayen (2007b), 'Predicting the dative alternation', in: G. Boume, I. Kraemer and J. Zwarts (eds.) *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science. 69-94.

Cruse, A. (2000), *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Deutschmann, M. (2003), *Apologising in British English*. Skrifter från moderna språk 10. Umeå: Institutionen för moderna språk, Umeå University.

Garzone, G. and F. Santulli (2004), 'What can corpus linguistics do for critical discourse analysis?', in: A. Partington, J. Morley and L. Haarman (eds.) *Corpora and Discourse*. Studies in Language and Communication 9. Bern: Peter Lang. 351-368.

Jucker, A.H., G. Schneider, I. Taavitsainen and B. Breustedt (2008), 'Fishing for compliments: precision and recall in corpus-linguistic compliment research', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam and Philadelphia: John Benjamins. 273-294.

Kepser, S. and M. Reis (eds.) (2005), *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter.

Kohnen, T. (2000), 'Explicit performatives in Old English: a corpus-based study of directives', *Journal of Historical Pragmatics*, 1(2): 301-321.

Kohnen, T. (2004), 'Methodological problems in corpus-based historical pragmatics. The case of English directives', in: K. Aijmer and B. Altenberg (eds.) *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*. Amsterdam: Rodopi. 237-247.

Kohnen, T. (2008), 'Tracing directives through text and time: towards a methodology of a corpus-based diachronic speech-act analysis', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam and Philadelphia: John Benjamins. 295-310.

Lüdeling, A. and M. Kytö (eds.) (2008), *Corpus Linguistics. An International Handbook*. Handbooks of Linguistics and Communication Science HSK. Berlin: Mouton de Gruyter.

McEnery, T., R. Xiao and Y. Tono (2006), *Corpus-based Language Studies. An Advanced Resource Book.* Routledge Applied Linguistics. London: Francis and Taylor.

Mey, J. (ed.) (2004), *Corpus Linguistics III*. Special issue of *Journal of Pragmatics*, 36(9).

Mukherjee, J. (2009), *Anglistische Korpuslinguistik. Eine Einführung*. Grundlagen der Anglistik und Amerikanistik 33. Berlin: Erich Schmidt.

O'Keeffe, A., M. McCarthy and R. Carter (2007), *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Partington, A., J. Morley and L. Haarman (eds.) (2004), *Corpora and Discourse*. Studies in Language and Communication 9. Bern: Peter Lang.

Romero-Trillo, J. (ed.) (2008a), *Pragmatics and Corpus Linguistics. A Mutualistic Entente*. Mouton Series in Pragmatics 2. Berlin: Mouton de Gruyter.

Romero-Trillo, J. (2008b), 'Introduction: pragmatics and corpus linguistics – a mutualistic entente', in: J. Romero-Trillo (ed.) *Pragmatics and Corpus Linguistics. A Mutualistic Entente*. Mouton Series in Pragmatics 2. Berlin: Mouton de Gruyter. 1-10.

Sanderson, T. (2008), *Corpus, Culture, Discourse*. Language in Performance 39. Tübingen: Gunter Narr.

Stenström, A.-B., G. Andersen and I.K. Hasund (2002), *Trends in Teenage Talk. Corpus Compilation, Analysis and Findings*. Studies in Corpus Linguistics 8. Amsterdam and Philadelphia: John Benjamins.

Stubbs, M. (1983), *Discourse Analysis. The Sociolinguistic Analysis of Natural Language*. Oxford: Blackwell.

Sunderland, J. (2004), *Gendered Discourses*. Basingstoke: Palgrave.

Teubert, W. and A. Čermáková (2004), *Corpus Linguistics. A Short Introduction*. London: Continuum.

Valkonen, P. (2008), 'Showing a little promise: identifying and retrieving explicit illocutionary acts from a corpus of written prose', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam and Philadelphia: John Benjamins. 247-272.

Verschueren, J. (1999), *Understanding Pragmatics*. Understanding Language Series. London: Arnold.

Vine, B. (2004), *Getting Things Done at Work. The Discourse of Power in Workplace Interaction*. Pragmatics & Beyond New Series 124. Amsterdam and Philadelphia: John Benjamins.

Wichmann, A. (2004), 'The intonation of *please*-requests: a corpus-based study', *Journal of Pragmatics*, 36: 1521-1549.

Part I

Pragmatics and discourse

Historical corpus pragmatics: Focus on speech acts and texts

Thomas Kohnen

University of Cologne

Abstract

This paper gives an overview of the major research in historical corpus pragmatics on speech acts and texts, adding new suggestions and insights. The section on speech acts deals with corpus-based diachronic descriptions of speech acts in the history of English and the methodological problem of retrieving speech acts in diachronic corpora. The section on texts contains a short overview of corpus-based descriptions of genres in the history of English and addresses the problems connected with the corpus-based analysis of functions and sections in texts. The final section, combining both perspectives, gives suggestions for a history of English as a history of genres and speech acts.

1. Introduction

Historical corpus pragmatics is, without doubt, still a very young field of study because it is the combination of two disciplines which are still very young themselves, corpus linguistics and historical pragmatics. It seems that the particularly 'adolescent' character of the combination stems from two facts. First, it was rather late that corpus linguists became interested in pragmatic topics. For example, Kennedy (1998: 174-180) offers roughly four pages of text on 'pragmatics and spoken discourse', and even a text book published in 2006 devotes little more than three pages to pragmatics in its survey section (McEnery, Xiao and Tono 2006: 104-108). So it does not come as a surprise that the historical part of pragmatics took a fairly long time to become a sub-field of corpus linguistics. Secondly, it appears that part of the seminal studies of historical pragmatics was in fact corpus-based and some of the roots of English historical pragmatics can actually be seen as recent developments in English corpus linguistics (see, for example, the early studies on the *Helsinki Corpus* which focus on pragmatic topics, Fries 1994; Kryk-Kastovsky 1995 or Taavitsainen 1995).

Given the adolescent character of the field, a survey of research may be instructive and risky at the same time. The following overview only covers slightly more than ten or fifteen years and the field may develop in new directions, making this account appear quite out of date in a few years. However, the focus on speech acts and texts may guarantee some continuity since central issues of historical pragmatics and corpus linguistics, which will probably remain

relevant during the next years, have been addressed under these headings. On the one hand, historical corpus pragmatics has contributed much to the exploration of uncharted territory in the history of English. Many, if not most, of the novel diachronic descriptions of speech acts and texts are corpus-based, and there is still a lot to be done to approach a 'pragmatic history of English'. On the other hand, historical corpus pragmatics has brought to light some of the thorniest methodological problems in the field and thus triggered innovative solutions to corpus-based methodology. Here, the discussion will probably go on during the coming years as well. Thus, the following survey, reflecting this situation, will have its chief emphasis on history and methodology.

The paper is structured as follows. After a brief look at definitions and distinctions there will be a section on speech acts and a section on texts, each containing a history and a methodology part. In the history part of the speech acts section, I will give a short overview of corpus-based diachronic descriptions of speech acts in the history of English; in the methodology part I will deal with the central problem of accessing and retrieving speech acts in a diachronic corpus. In the section on texts, the history part will contain a short overview of corpus-based descriptions of genres in the history of English; in the methodology part I will deal with the problem of the corpus-based analysis of functions and sections in texts. In the concluding chapter both perspectives will be combined. I will address some challenges and make some suggestions for a history of English as a history of genres and speech acts.

2. Definitions and distinctions

What is historical corpus pragmatics? We may start with a brief definition of historical pragmatics and then move on to the qualification in terms of corpora. The beginning of historical pragmatics as a proper field of study can be linked to Jucker's collection of case studies (Jucker 1995) and the survey and outline of research perspectives included there (Jacobs and Jucker 1995). According to this approach, historical pragmatics can be divided up into 'pragmaphilology' and 'diachronic pragmatics', two terms which have become fairly established. Pragmaphilology comprises the pragmatic analysis of historical data belonging to one particular period (for example, the study of English sermons stemming from the fifteenth century, of politeness in Old English, or of address terms in Shakespeare). Diachronic pragmatics deals with the diachronic development of pragmatic phenomena across different historical stages (for example, the development of the different functions of *thou* and *you* in the history of English or the development of directive speech acts in the history of English).

There is a different, that is, more focussed approach to historical pragmatics, which was introduced by Traugott. She sees historical pragmatics as "a usage-based approach to language change" (2004: 538). This approach is more focussed in that it concentrates on work on semantic change and lexicalisation, in

particular, the "discourse pragmatic origins and motivations for semantic and lexical change" (2004: 539).

Brinton (2001), surveying the related field of historical discourse analysis, provides a third way of looking at historical pragmatics, which, however, is in part similar to the approach suggested by Jucker. She introduces a threefold distinction. Historical discourse analysis proper is the strictly synchronic study of discourse forms and functions in earlier periods of a language. This corresponds to Jucker's pragmaphilology. Diachronically oriented discourse analysis comprises the study of discourse structures and functions over time. This is obviously quite similar to Jucker's diachronic pragmatics. Discourse-oriented historical linguistics is the application of discourse analysis to historical linguistics, in other words, the study of discourse-pragmatic factors and discourse motivations in language change. This seems to me to be in part analogous to Traugott's approach (which, as was pointed out above, is more focussed).

Thus, summing up, one can say that in historical pragmatics we are basically dealing with three aspects: the synchronic analysis of pragmatic phenomena in a historical period, the diachronic analysis of pragmatic phenomena across several periods and the pragmatic motivations for language change. Such a subdivision may be manageable and clear enough, but, with regard to the diversity of topics dealt with under the heading of historical pragmatics, the opposite situation seems to prevail (on this variable situation see, for example, Brinton 2001: 138-139 and Taavitsainen and Fitzmaurice 2007: 12-13). One could claim that the inventory of topics studied in historical pragmatics is as multifarious as the items on sale in a big department store. This threatens to blur the neat subdivisions introduced above.

Jucker (2008) has recently come up with a useful classification which may help in putting some order into this variety. Jucker subdivides the field according to the (pragmatic) units of analysis. The lowest level, expressions, includes, for example, interjections, discourse markers and address terms. The next level comprises utterances, which are typically represented by speech acts. At the third level we find discourses and genres, that is, whole texts and groups of texts. The last level includes discourse domains, that is, the socially defined institutions and frameworks for the formulation and dissemination of texts (for example, religious discourse, discourse of science, discourse of mass media). This taxonomy is extremely useful, but it leaves out some cross-sectional aspects which touch upon several units, like implicature, politeness, orality and literacy and others. For example, aspects of politeness are relevant at the level of expressions (address terms), utterances (speech acts) and discourses/genres. However, it does not seem to be difficult to supplement Jucker's classifications with such cross-sectional aspects.

Now, what is historical corpus pragmatics? Basically, one could say that all studies in the areas and on the topics of historical pragmatics as pointed out above belong to historical corpus pragmatics in so far as they use corpora, that is, in so far as they are corpus-based. The question is, of course, what one should choose to call 'corpus-based'. Does it imply the use of any corpora or any data in

any combination or a more systematic use of corpora, involving the principles of corpus-linguistic analysis (as regards corpus size, representativeness, balance etc.)? I prefer a narrower sense of the term corpus-based and thus a narrower sense of historical corpus pragmatics. Some studies (especially in the field of discourse-oriented historical linguistics, for example, Traugott 2004) seem to employ as many data (or corpora) as possible in order to retrieve as many (early) relevant items as possible. A narrower concept of historical corpus linguistics would not include such studies, but it is sometimes difficult to draw a clear line. If one chooses the narrower notion of historical corpus pragmatics, it seems that a large part of the field is covered by studies in the synchronic analysis of pragmatic phenomena in a historical period or the diachronic analysis of pragmatic phenomena across past periods. Speech acts and texts are central topics in these areas.

3. Speech acts

3.1 History: corpus-based descriptions of speech acts in the history of English

As was pointed out in the introduction, this chapter starts with corpus-based accounts of speech acts in the history of the English language and will then go on to problems of methodology.

3.1.1 Case studies in illocutionary development

Corpus-based work on the history of English speech acts has been very productive during the past few years. One of the first book-length accounts of illocutionary histories was Arnovick (1999), a work which traces the development of selected English speech acts and speech events (for example, flyting/sounding, promising, cursing, blessing, and greeting). A good example of a corpus-based analysis is the history of *Good-bye*, which, as a contracted form, goes back to the phrase *God be with you* and which turns from an explicit blessing with a possibly implied greeting to a polite form of ending a conversation. Using the English Drama collection of the *Chadwyck-Healey "Literature Online" Database*, Arnovick demonstrates that most of the Early Modern English occurrences of *God be with you* must be seen as blessings, with the simultaneous function of a parting greeting. At the beginning of the eighteenth century, when contraction is much advanced, blessings dwindle away, leaving the polite terminal utterance as the only function of *Good-bye*. Arnovick calls this development, whereby one specific illocutionary function is turned into a structural discursive function, 'discursisation'.

Arnovick's book is certainly a pioneering work, but, seen against the background of a strictly corpus-based analysis, it turns out that not all her case studies are really corpus-based but in some cases rely on fairly restricted data (for

example, in the case of curses). A second problem is the fact that she only includes a very restricted set of different manifestations of the speech acts under discussion, that is, in many cases only one manifestation (for example, *Good bye* as the only manifestation of greeting). But despite these restrictions, Arnovick (1999) is an innovative work, which should be included in the field of historical corpus pragmatics.

3.1.2 Tracing speech acts in pragmatic space: expressives

Another important, largely corpus-based approach to the history of English speech acts is Jucker and Taavitsainen's account of expressives. The authors trace expressives in what they call the multidimensional pragmatic space. In their first study (Jucker and Taavitsainen 2000) they deal with insults, while recent investigations are about compliments and apologies (Taavitsainen and Jucker 2008; Jucker and Taavitsainen 2008).

Jucker and Taavitsainen (2000), when looking at insults in the history of English, point to the difficulty of drawing a continuous line of development connecting different manifestations of the speech act over time. They mention, for example, flyting (boasts and challenges stemming from the Old English heroic tradition), sounding (ritual insults associated with urban African-American adolescents) and flaming (insulting conduct found in news groups on the internet). These are fairly disparate realisations of insults, showing differences in terms of time, culture and linguistic means. In order to capture this diversity, Jucker and Taavitsainen suggest a diachronic analysis of speech acts in a multidimensional pragmatic space, which is, in this case, the pragmatic space of antagonistic behaviour. Here the individual manifestations of insults can be seen in the specific settings of their culture and period and in relation to neighbouring speech acts. The account is based on a minimal obligatory definition of an antagonistic act ('a predication about the addressee which is perceived as disparaging'), which can then be developed in terms of the different parameters of the pragmatic space, for example, ritual vs. creative (that is, whether the order of elements of the insult follows a predictable pattern or not), conventional vs. particular (that is, whether the insult is usually taken as an insult by all members of the speech community or not), ludic vs. aggressive (that is, whether the insult is only playful or followed by real violence), and so on. Using this pattern, Jucker and Taavitsainen analyse snapshots of different realisations of insults in their specific culture and period from Old English to Present-day English. Here again, one can argue that in this study their data collection cannot be called strictly corpus-based since they only pick out examples from reading and the relevant literature. In fact, the major aim of this study was not a corpus-based representative analysis but rather the exploration of the concept of 'pragmatic space'. But in this it laid the foundation for further corpus-based studies.

The initial study of insults was supplemented by a strictly corpus-based investigation (Taavitsainen and Jucker 2007) in which the authors studied the set of speech-act verbs of verbal aggression, using the *Helsinki Corpus*, ARCHER, the *Corpus of Early English Correspondence Sampler* and the *Chadwyk-Healey*

drama and fiction corpora. The corpus searches showed that the speech-act verbs were mostly used descriptively, that is, not as performatives. The descriptive uses reveal the changing perceptions of insults and what constituted insults during the history of the English language. Thus, they give an instructive ethnographic picture of the pragmatic space of verbal aggression. This picture starts with descriptions of God's anger and only later shifts to descriptions and negotiations of antagonistic behaviour in conversation. Typical phrases are *Do you mock me? Do not mock me!* It also seems that the range of topics which are felt to be insulting becomes broader in the course of time. They include name-calling, remarks on family relations and money, personal skills and, of course, national stereotypes. Examples (1) and (2) illustrate family relations and national stereotypes, respectively.

(1) "You ask a strange question, sir."
"Which I will answer for you in the negative," said Luther. "You know neither yourself nor your parents."
"Do you wish to insult me?" cried the other, reddening and somewhat confused.
(*A Tale of the Little Miami* [1848], p. 46, cited in Taavitsainen and Jucker 2007: 132)

(2) Now this was every Word of it False, and was only form'd by his nimble Invention, to insult us as English-men;
(Defoe, *Colonel Jack* [1723], p. 128, cited in Taavitsainen and Jucker 2007: 134)

The ethnographic approach was also used in a follow-up study on compliments (Taavitsainen and Jucker 2008), where typical speech-act designations of compliments were employed to retrieve the speech acts in the *Chadwyk-Healey* collections. Here Taavitsainen and Jucker point to the prevailing interpersonal functions of compliments and the changing topics used in compliments. For example, looks are prevalent in earlier texts, but not possessions and taste.

Another follow-up study (Jucker and Taavitsainen 2008) focussed on apologies. Here the authors searched for lexicalised and routinised formulae used in apologies (for example, *excuse me*, *pardon me*, *I beg your pardon*, *I am sorry*, *forgive me*, *I am afraid*) in the *Chadwyck Healey On-line Corpus* (1500-1660). Among the interesting results were that Early Modern English apologies seem to have been less routinised and more explicit and that they involve the addressee's forgiveness rather than the addressor's regret. Thus, people in the Early Modern period would typically prefer *forgive me* to *I'm sorry*, a realisation which, on the surface, requires the addressee to perform an act (instead of showing the addressor's state of mind).

3.1.3 Directives: politeness and indirectness

Another important class of speech acts in the history of English are directives. Corpus-based investigations of directive speech acts have been a major focus of

my own research during the past years and prominent issues have been politeness and indirectness.[1]

Instead of using the framework of pragmatic space, I started with the very general Searlian definition of directives as the class of speech acts which count as an attempt by a speaker to get the hearer to carry out an act (Searle 1976: 11). The interesting point about directives is that they are usually seen as face-threatening acts since they can violate the addressee's negative face, and that in English indirect manifestations are used as face-saving strategies.[2] The question then is: When and how did indirect or more polite manifestations of directives evolve in the history of English? Against the background of contemporary English, the following four manifestations of directives typically imply different assumptions about the relative politeness or directness of the speech act, with the explicit performative (example (3)) at the less polite end and the interrogative (example (6)) at the more indirect end of a scale:

(3) I order you to leave.
(4) Go!
(5) I would like you to leave.
(6) Could you please leave?

Was the complete range of manifestations of directives already available in earlier periods of the English language (for example, in Old and Middle English) or did they gradually develop in the course of time? In order to find an initial answer to this question, my first studies focussed on specific manifestations of directives in multi-genre corpora (for example, directive performatives or imperatives in the *Helsinki Corpus*), whereas later on I focussed on all manifestations of directives in a genre-based corpus (for example, all directives in a corpus of sermons or prayers).

The general trends several studies seem to support are that direct manifestations decrease during the history of the English language, while face-saving manifestations slowly evolve during the Early Modern period. However, it seems that some qualifications are necessary here. These involve the question of whether politeness has always been linked to indirectness and, secondly, whether the requirements of politeness may be suspended by genre conventions.

In the present section I will give a short overview of three investigations (Kohnen 2000a, 2002, 2004) focussing on specific manifestations, whereas the genre-based studies will be mentioned in the following section, which is devoted to methodological issues.

Kohnen (2000a) is an analysis of directive performatives in the Old English section of the *Helsinki Corpus* and in the LOB corpus. Here the most noteworthy result was that the frequency of directive performatives in Old English is more than seven times as high as that found in the LOB corpus (4 vs. 0.55 per 10,000 words). In addition, with regard to the different classes of directive speech-act verbs, the frequency of the verbs denoting asking and ordering in the *Old English Corpus* is far higher than in the LOB corpus (2.5 vs.

0.14 and 1.5 vs. 0.07). By contrast, verbs denoting suggestion or advice (where it is left to the addressee to decide which course of action to take) are predominant in the LOB corpus, whereas the *Old English Corpus* contains none. Thus, the data suggest that the frequency of performative directives decreases in the history of the English language. In addition, the large number of 'suggest/advice verbs' in the LOB corpus indicates that, if performatives are used today, they are not used in face-threatening acts.

Kohnen (2004) deals with another prominent direct realisation of directives, the imperative. I looked at imperatives in the religious treatises in the tagged and enlarged version of the *Helsinki Corpus* (PPCME2) and in the *Brown* and LOB corpora. Here the data again suggest a significant drop in frequency. Whereas the frequency of imperatives between 1200 and 1375 is fairly high (4.5 in 1,000 words), it is only 1.9 in Late Middle English and 0.6 in the LOB corpus and 0.83 in the *Brown Corpus*. A picture similar to the development of the performatives emerges. Another manifestation which is usually called direct seems to decrease in the course of the history of English.

Kohnen (2002) looked at the development of two typical indirect manifestations of directives in the *Helsinki Corpus*, the speaker-based declaratives expressing speaker volition (*I would like you to leave.*) and hearer-based interrogatives questioning the volition/ability of the addressee (*Could you please leave?*). Here the data showed that speaker-based manifestations develop slowly during the sixteenth and the first half of the seventeenth century, whereas hearer-based interrogatives do not really spread until the end of the seventeenth century. In all, the general frequency of both constructions is still rather low at the end of the Early Modern period (1.1 and 0.82 in 10,000 words).

So the general picture suggested by this short overview is that so-called face-saving directives did not develop until Early Modern English, while the number of those manifestations which may be called direct decreased. The late advent of indirect directives was supported by a very recent study of conventional indirectness. Culpeper and Archer (2008) looked at the manifestations of requests in their socio-pragmatic corpus of trials and plays dating from 1640 to 1760. In their analysis (which was based on the so-called *Cross-cultural Speech Act Realisation Project*, see Blum-Kulka et al. 1989) they found that requests in their data are quite different from those found to be prevalent in modern languages, with a low proportion of conventional indirectness and a large share of so-called impositives (that is, imperatives, performatives, hedged performatives, obligation- and want-statements). On the one hand, this would support the late evolution of indirect speech acts, especially of the interrogative hearer-based kind. On the other hand, the authors argue that lack of indirectness may not necessarily imply lack of politeness, since the correlation of directness with less politeness may in fact be culture-specific and is not borne out by the data.

This raises the general question of whether the picture of directives becoming less explicit, less direct and less face-threatening can remain unqualified. It may be questioned from the perspective of the cultural conventions of politeness, but also from the perspective of a corpus-based methodology. Here

there are quite a few issues which call for a re-assessment of the general, though tentative, conclusions drawn from the initial studies. These methodological problems are the topic of the next section.

3.2 Methodology: tracing speech acts through text and time

The methodological problems of a diachronic speech-act analysis have been mentioned quite early in the development of historical pragmatics (see, for example, Jucker 2000 or Bertuccelli Papi 2000). One of the advantages of corpus-based studies is that they have brought these and similar difficulties quite clearly to the surface. In my view, the most important problems which are typical of a corpus-based diachronic speech-act analysis can be summed up in three points.

The first point refers to the well-known fact that the relationship between form and function in speech acts is hardly fixed, and different manifestations are unpredictable, especially in past periods of a language. Researchers have to put up with a more or less eclectic analysis. They can either collect typical illustrative realisations in different periods of the English language ("illustrative eclecticism", see Kohnen 2004); or they can start with a selection of typical patterns which are traced by way of a corpus-based analysis throughout the history of English ("structural eclecticism", see Kohnen 2004).

The second problem relates to the fact that we cannot locate the development of a specific manifestation of a speech act within the larger picture of all manifestations of that speech act in a corpus. Since we cannot cover all the manifestations, there is always a residue of "hidden manifestations" (Kohnen 2007). For example, does the tendency towards indirectness noted in Early Modern English directives suggest that the more face-threatening manifestations are replaced by more polite ones or that the indirect realisations are used in addition to the less polite directives? The underlying quantitative and qualitative proportions are not revealed and so the picture is fragmentary.

Thirdly, there is the risk of mixing genres and registers in the analysis. When researchers use a multi-genre corpus, they may, even if the data should include a relatively full list of realisations, mix up directives employed on different occasions and in different genres. For example, decreasing frequencies in treatises and increasing frequencies in letters during the same period could combine and simulate a stable situation, which in fact does not prevail.

There are several ways in which researchers can approach and tackle these problems. Most suggestions try to deal with what I call the problem of 'unpredictable manifestations', the first point mentioned above.

3.2.1 Speech-act verbs and the ways they give access to speech acts in corpora

One way researchers deal with these problems relies on speech-act verbs and the ways they give access to speech acts in corpus data. As was shown above, Taavitsainen and Jucker (2007) use speech-act verbs of verbal aggression in order to access insults in a large corpus. The set of speech-act verbs of verbal aggres-

sion can be determined for the major periods in the history of English and can thus be systematically retrieved in a historical corpus. This approach can cover all the items in a historical corpus where speech-act verbs of a certain kind are mentioned and can thus, as was shown above, provide a highly instructive ethnographic picture of the respective speech acts. It can also serve to access explicit performative manifestations of speech acts (see Kohnen 2000a, 2000b). But it cannot cover all the manifestations of a specific speech act in a historical corpus.

3.2.2 Fixed, recurring manifestations of speech acts

Another approach relies on the fact that many speech acts are realised by more or less fixed, recurring patterns. A corpus-based analysis can start with such common patterns of linguistic expressions, convert them to more abstract search strings approximating them and then test their precision and recall in large corpora, that is, test whether they achieve correct identification and comprehensive coverage of the speech act under investigation.

For example, Jucker, Schneider, Taavitsainen and Breustedt (2008), while "fishing for compliments", look at linguistic patterns which are typically used as compliments and work out approximations to the number of compliments in the *British National Corpus*. Valkonen (2008) investigates explicit performatives containing speech-act verbs of promising. One of his interesting results is that 97 percent of all instances in his large corpus are based on the verbs *promise*, *swear* and *vow*.

These and similar studies certainly yield instructive results, but their reliability is based on the extent to which a speech act is realised by fixed and routinised expressions. Since they always begin with a formal specification of the speech act under investigation, they risk missing the more creative, indirect or simply uncommon manifestations and thus cannot solve the problem of the "hidden manifestations".

3.2.3 A genre-based empirical bottom-up methodology

In a more or less complementary approach to the studies mentioned in the previous sections, I suggested a genre-based bottom-up methodology (for a more detailed account of this methodology, see Kohnen 2008). This methodology starts with relatively small diachronic pilot corpora of selected genres (for example, sermons, letters, prayers etc.) and analyses all the manifestations of a particular speech act (for example, directives) in these pilot corpora. This analysis necessarily proceeds 'by hand', that is, all the text excerpts have to be read, considering carefully which sections of text might serve the function of the relevant speech act. This microanalysis will produce a preliminary inventory of the manifestations of the speech act under investigation. With more genres, the list of different realisations will probably grow. However, it seems reasonable to assume that the more genres included, the less 'new' manifestations will be found. In a final step, selected manifestations and their distribution can be tested

in larger multi-genre corpora in order to assess the frequency and distribution of the various manifestations in a more comprehensive setting of language use.

Given an appropriate number of genres and pilot corpora, such a methodology should produce a fairly detailed inventory of the different manifestations of a speech act across time, approaching a reasonable level of completeness and representativeness. This, it is hoped, should significantly increase the retrievability of speech acts in diachronic corpus-based studies, thus reducing the impact of the problem of unpredictable and hidden manifestations. With regard to the third problem mentioned above (mixing genres and registers), this method will also enhance our knowledge about the distribution of speech acts and their different manifestations across genres.

So far, I have tested the method only in pilot studies on directive speech acts in selected genres, but all the studies clearly suggested two major results. First, there is a large majority of common manifestations of directive speech acts in the history of English and, secondly, there are also clear genre-specific profiles in the frequency and distribution of directives. The following four examples are used to illustrate these results (tables 1 to 4).

Table 1 below shows the distribution of the manifestations of directives in a study of sermons, letters and prayers in a period covering Old English, Late Middle English, Early Modern English and the late twentieth century. The largest proportions are covered by imperatives, modals and performatives. The proportion of the manifestations whose different orthographic realisations are difficult to determine in principle (mainly the indirect manifestations and some modal expressions) does not go beyond 12 percent in these data.

Table 1: Distribution of manifestations of directives in sermons, letters and prayers (%)[3]

	10th/11th c.	15th c.	16th c.	17th c.	late 20th c.
performatives	4	22	15	11	22
1st person imp.	32	7	2	1	7
2nd person imp.	14	35	71	78	47.5
3rd person imp.	11	0	1	1	0
modals	40	28	6	3	15.5
indirect	0	8	5	6	8

It also seems that the three genres examined form more or less clearly discernable patterns in the distribution of directives. Table 2 below shows that sermons provide a mixed picture of modals, second-person and first-person imperatives, with a growing proportion of indirect speech acts in the last period.

Seen against the background of politeness, sermons do not form a coherent picture. First-person imperatives, which are usually seen as strategies of positive politeness, have a large share in Old English (32 percent); indirect manifestations are missing in the seventeenth century and there is a relatively large proportion of second-person imperatives in the last period (30 percent). All these are indica-

tions which do not conform to a smooth uninterrupted growth of 'face-saving' and a decrease in 'direct' manifestations. Here it seems that genre-specific features have to be taken into account as well.

Table 2: Distribution of manifestations of directives in sermons (%)

	10th/11th c.	15th c.	16th c.	17th c.	late 20th c.
performatives	4	5	1	3	0
1st person imp.	32	16	11	15	18
2nd person imp.	14	25	43	33	30
3rd person imp.	11	0	8	14	0
modals	40	53	33	36	39
indirect	0	1	5	0	13

Table 3 below describes the situation in private letters, which seems to be rather different from sermons. There is a clear predominance of second-person imperatives (ranging between 38 and 60 percent), which, from the fifteenth through the seventeenth centuries, are accompanied by a large proportion of performatives. The proportion of performatives, however, decreases in the seventeenth century and in the last period no performative is found. The other type of manifestation which plays a major role in letters is indirect directives. They show a relatively high proportion from the beginning, which grows significantly, so that indirect directives are the second major type of directives in the late twentieth century. Thus, letters seem to follow the pattern of development suggested by earlier research: a decreasing share of performatives and an increasing proportion of indirect manifestations.

Table 3: Distribution of manifestations of directives in letters (%)

	15th c.	16th c.	17th c.	late 20th c.
performatives	37	37	26	0
1st person imp.	0	2	1	0
2nd person imp.	43	38	44	60
3rd person imp.	0	0	0	0
modals	6	5	5	4
indirect	14	18	24	36

The situation in prayers seems to be quite simple (see table 4, below). There is a great predominance of second-person imperatives (which is slightly lower in the last period). There are also some performatives. This situation does not change very much over the centuries, and it seems likely that considerations of politeness had no effect on the selection of the manifestations of directives.

Table 4: Distribution of manifestations of directives in prayers (%)

	16th c.	17th c.	late 20th c.
performatives	13	8	41
1st person imp.	0	0	0
2nd person imp.	84.5	91	59
3rd person imp.	0.4	0	0
modals	0.2	0	0
indirect	2	1	0

In summary then, it appears that the three genres examined form more or less clearly discernable profiles in the distribution of directives. It also appears that the general trend towards more polite manifestations must be qualified against the background of genre-specific requirements. Politeness is certainly an important factor in some, but not all, genres. In addition, the data suggest that speech acts should be accessed in the contexts of the texts and genres in which they occur. This brings us to the next section.

4. Texts and genres

The diachronic study of texts and genres seems to be a logical continuation of the study of speech acts. Speech acts of past centuries are usually preserved in written texts, and texts of past centuries can thus be seen as complex constellations of speech acts or, rather, text acts. This, of course, involves the same methodological problems which we met in the diachronic analysis of speech acts. But texts and text acts clearly go beyond speech acts in that they also exhibit larger and more complex sections which correspond to more general text functions. Such higher-level text functions are super-ordinate to individual text acts because they may comprise several text acts. For example, the petition section of a prayer may comprise several directive acts. But these higher-level text functions are subordinate to the overall purpose of the whole text. For example, a petition is only one typical element of prayers, besides invocation, adoration and others (on these see below). Categories like genre and text type are of course even more complex entities, which clearly go beyond the category of speech act.

In what follows I will first give a short overview of corpus-based research on some major genres in the history of English and then go on to some methodological problems relating to super-ordinate text functions.

4.1 History: corpus-based descriptions of genres in the history of English

During the past few years quite a few corpus-based studies have been published which have contributed to the description of individual genres and their development in the history of English. Many of them are linked to recent work on genre-based corpus projects. I will briefly mention some quite prominent examples.

Several descriptions of the development of the genre of letters have grown out of the work on the *Corpus of Early English Correspondence* (Nevalainen and Raumolin-Brunberg 1996; Nevala 2003; Nevalainen and Tanskanen 2004). Work on the evolution of the genres connected with medical prose and academic writing has been done in connection with the corpus of *Middle English Medical Texts* (MEMT) and its Early Modern English counterpart (Taavitsainen and Pahta 2004; Taavitsainen 2005). Studies linked to MEMT focus on the stage of the vernacularisation of scientific writing during the Middle English period, when treatises associated with medical writing formed a major force in this process, establishing a rich inventory of (mainly new) English subgenres.

Work on pamphlets, in particular with regard to questions of text type and genre variation, has been discussed in connection with the *Lampeter Corpus of Early Modern English Tracts* (Schmied and Claridge 1997; Claridge 2000, 2001). The *Zurich English Newspaper Corpus* has given rise to investigations on the evolution of English newspaper and media discourse, in particular, on text classes and subgenres, their structure and their development (Fries 1997, 2001, 2002; Jucker 2005). Genres of spoken language (trials, depositions, dialogues in fiction and handbooks) have been described in connection with the compilation of the *Corpus of English Dialogues* (Archer 2005; Culpeper and Kytö forthcoming).

There are also investigations which are based on diachronic multi-genre corpora (here, above all, on the *Helsinki Corpus* and ARCHER). Finegan and Biber (1989) and Biber and Finegan (1992) made comparative diachronic analyses of three written genres (essays, fiction, and personal letters) and two speech-based genres (dialogues from plays and from fiction) from the seventeenth century up to 1950. Here the focus was on the relative 'orality' and 'literacy' of the genres. In my own work on the development of participle constructions (Kohnen 1997a, 1997b, 2001), I linked the spread of some participle constructions with developments in text types, pointing out some of the more noticeable changes in them (for example: in narrative prose a change from small fragmentary anecdotes to a larger and more coherent narrative framework; in chronicles a change from fragmentary listings to the continuous report and coherent description of the more modern accounts; or in petitions/statutes a tendency from large narrative passages to more formal accounts of regulations).

This short overview of corpus-based descriptions of genres in the history of English is not meant to give a comprehensive picture, but it may reveal that we are basically dealing with separate descriptions of individual genres or sub-genres belonging to a discourse domain (like medical discourse or news discourse). Within these 'individual snapshots' we may miss an underlying coherent network linking the diversity of the different genres. In addition, many descriptions often rely on a restricted set of linguistic features (for example, participle constructions, address terms etc.).

I would claim that a more coherent account can be achieved if we include the higher-level functions and sections of texts and the functional profiles of genres in the diachronic descriptions of genres. These higher-level functions may capture the common elements which link several genres (for example, exposition

and narration sections can be found in several genres); on the other hand, their specific combination in the functional profile of the genre and their 'embeddedness' in a specific discourse domain provide the individual character of the genres (for example, prayers and private letters may share several functions (address, petition), but they differ in their specific combination and the respective discourse worlds). If we want to access the higher-level functions and sections in texts, using corpus-based tools, we face serious methodological problems. These will be addressed in the following paragraphs.

4.2 Methodology: approaching functions and sections in texts

Against a corpus-based background, the important question seems to be: How can we access the more complex structures and larger units associated with functions and sections in texts, using corpus-based tools and methods? Here I would again opt for a traditional, 'philological' approach which starts with a manual analysis of a small genre-based pilot corpus. However, the analysis should be top-down and will be determined by the functional profile of the genre. That is, based on our knowledge of the functional profile of the genre and its respective domain, the basic text functions linked to text sections are specified in advance.

For example, a prayer, in the domain of religious discourse, can be seen as an address or a message to God (or a saint) which usually includes invocation (the act of calling upon God or the saint), petition (the various requests addressed to God), thanksgiving, confession/profession (confessing one's sins and professing one's faith) and adoration (acts of praising and worshipping). Another example: a sermon is a piece of religious instruction with a priest (or a person in a similar clerical function) addressing a religious congregation. It typically includes the major text functions associated with religious instruction: exhortation, exposition, narration, exegesis, and sometimes argumentation (for a more detailed account of religious instruction, see Kohnen 2007).

Thus, the top-down analysis is complementary to the bottom-up analysis suggested in 3.2.3 above. The top-down analysis provides the functionally determined text sections as a general framework for the bottom-up analysis, which provides a detailed microanalysis of these sections.

There is a major recent corpus-based approach to contemporary discourse structure (Biber, Connor and Upton 2007) which uses a somewhat similar methodology and terminology. The top-down analysis suggested here is quite similar to what Biber, Connor and Upton call a genre-based "move analysis" (on move analysis see also Swales 1981 and 1990). The specific communicative functions of a genre, that is, its 'moves' or 'move types', are determined in advance and then the respective stretches of texts linked to these moves are analysed. It seems, however, that in this approach the moves are fairly specialised and quite genre-specific in their functions (compare, for example, 'identifying a gap in previous research' as a move type in the genre of research article with the general text function of 'exposition' in the genre of sermon above; on general text functions like exposition see Werlich 1976). Biber, Connor and Upton's bottom-up analysis, on the other hand, is quite different from the bottom-up analysis

suggested here, in that it is an automatic computerised segmentation which is based on the distribution of the vocabulary.

Within the approach suggested here, a first step is a manual coding which delimits the sections covered by the identified text functions in the texts of the pilot corpus. In a second step, the structure of these sections is analysed with regard to the typical combination of text acts and their typical manifestations. For example, an analysis of prayers could show whether the petition section in prayers typically contains directive acts and to what extent they are realised by imperatives, performatives or other manifestations. Such an analysis would also include typical boundary features, items which typically delimit or introduce new text sections. Thus, the initial micro-analysis of a pilot corpus would suggest a preliminary idea of the most salient morpho-syntactic and pragmatic features of the functionally defined text sections. For example, my analysis of a pilot corpus of Middle English and Early Modern English prayers has revealed the following typical features. The invocation section typically contains designations for God (for example, *Lord*, *Jesus* etc.), the particle *oh*, appositions and/or relative clauses. The petition section comprises mostly explicit performatives and imperatives. For thanksgiving, confession and adoration we find mostly explicit performatives as well. Boundary features can be found in the particle *oh* or typical ending formulae (e.g. *amen* etc.). Many of these features can be found in the short text passage below (italics added), which contains a Late Middle English prayer to the trinity.[4]

(7) [invocation] *O BLESSYD trinyte, Fader, sone, and holy ghoost*: thre persones, and one god,
[confession/profession] *I byleue* with my herte, and *confesse* with my mouth, al that holy chirche bileueth and haldeth of thee: .. and *I proteste* here before thi maieste, that I wyl lyue and deye in thys fayth, <P 285> and continue al my lyf. And *I knowlech* thee, my good fader and maker of al the world, and me thy poure creature, subgette, and seruaunte. And *make* to the *faith and homaige* of my body and of my soule, ..
[thanksgiving] and wyth al my hert *I remercye and thanke thee*. And in the sygne of recognisaunce and knowleche, I paye to thee thys lytel tribute on the mornyng, and on the euenyng thys:
[adoration] that *I adoure and worship thee* with hert and mouth, in feith, in hope and in charite, wyth thys lytel oryson and prayer, whyche alle apperteyneth to thy blessid maieste, seignorie and diuinite,
[petition] and *humbly require thee* of thre thynges. The *fyrst*, is mercy and forgyueness of as many euylles and vyllain synnes, as I haue doon and commysed in thyme passed agenst thy wylle. The *secounde*, plese thee to gyue me grace, that I may serue thee and fulfille thy commaundementis, withoute to renue ne to falle in to dedly synne. The *thyrd* is, that at my deth and at my gret nede thou wylt socoure me: and gyue me grace that I may haue remembraunce of the blessid passion, and contricion of my

> synnes: and that I may deye in thyn holy faith, and finably may com to the glorie eternal, wyth alle the sainctes of heuen. *Amen*.
> (Prymer of Salisbury, Maskell, 1882: 284-285)

In the above example, the invocation section comprises the particle "o", designations for God ("blessyd trinyte, fader, sone, and holy ghoost" etc.), partly arranged in appositions. The confession/profession section contains mainly performative expressions which make explicit that the person praying conforms to the correct Catholic belief ("confesse", "proteste", "knowlech"). The next section comprises thanksgiving, syntactically linked to adoration, which again shows performative expressions arranged mostly in doublets ("remercye and thanke", "adoure and worship"). The petition section is also introduced with an explicit performative, and then takes the form of an enumeration marked with numbers. "O" and "Amen" can be seen as boundary markers.

Once a preliminary pattern is established, this linguistic profile of the functional text segments can be used to track similar text segments in larger corpora of the same genre. For example, in prayers the respective explicit performatives of the petition, thanksgiving, confession and praising sections can easily be used to trace the respective sections in larger prayer corpora. Boundary markers like *oh* and *amen* may be used for the same purpose. Thus, with the combination of the top-down and bottom-up analysis we may get an initial tool for accessing larger, functionally defined text segments with corpus-based methods. Of course, this method only works to the extent that the texts and their sections exhibit the pattern and features of the linguistic profile. But my experience with the prayer texts has been quite encouraging so far, since the respective sections can easily be accessed using the performatives and other typical features.

On a higher level, the level of the genre, this analysis can, in addition, give a first idea of the individual proportions of the functionally defined text sections, and it gives a first indication of the general linguistic profile of the genre, that is, its salient morpho-syntactic and pragmatic features. For example, salient features of prayers seem to be performatives, first- and second-person pronouns (many of which come along with the performative expressions), address terms and appositions. The important point for a corpus-based analysis is that the features of the linguistic profile refer to one genre, and the typical manifestations of its core sub-functions, not to a pre-selected set of morpho-syntactic features which are applied to all texts of all genres in a corpus. The salient features of prayers, for example, might have easily been lost sight of if we were to use a general set of morpho-syntactic features for the analysis (for example, the features employed in Biber's multi-feature analysis, which do not comprise performatives, appositions and address terms; for a recent version of this list see Biber, Connor and Upton 2007: 267-271).

The salient features contained in the linguistic profile of a genre may be used in larger genre-based corpora to trace changes in the linguistic profile across centuries and to compare different genres. For example, prayers seem to be a

fairly stable genre as many salient features are preserved across the centuries (see also table 4 with the distribution of the manifestations of directives in prayers). But if a decrease in explicit performatives (for example, of verbs of thanking) were noted, this would indicate a major change in the genre. However, the corpus data indicate that this does not seem to have happened across the centuries.

5. Conclusion: the history of English as a history of genres and speech acts

One major project in the field of historical corpus pragmatics, where the two methods sketched out in this article could be combined, is a history of English which is devised as a history of genres and speech acts. So far, a coherent research agenda which might specify the individual steps towards such an aim does not seem to exist. This survey article can hardly provide a fully-fledged program either, but it may sum up some major challenges and indicate some possible ways of proceeding.

The first major question which needs to be answered is: Should such a history give an account of all the major genres and speech acts found in the history of English following the pattern of the traditional periods or should it depict separate 'case histories' of individual genres/items? In other words, should its major orientation be pragmaphilology or diachronic pragmatics? I would advocate a pragmaphilological approach, that is, a combination of (synchronic) analyses of the traditional periods, that is, Old English, (early and late) Middle English, Early Modern English, Late Modern English and Present-day English, as a starting point. Apart from the fact that such a design would fit the established image of a 'history' of the English language, it could also provide a basic framework where the position and impact of individual case histories could be located.

Once this general question is solved, the next challenge is to provide an inventory of the genres and speech acts in the major periods in the history of the English language. It goes without saying that there is no such thing as a complete and generally accepted list of genres and speech acts covering all the sections of language use in English across the centuries. Given the different aims and orientations of pragmatic and text-linguistic approaches, in addition to the basic difficulties in dealing with historical data, such an inventory is not likely to appear. However, we can approach a limited list of speech acts and genres.

With speech acts it seems reasonable to start with the major classes of speech acts and access them through texts and genres. Speech acts are less complex than texts and genres (if we follow the account of traditional speech-act theory) and there are fairly comprehensive classifications of speech acts (for example, Searle 1976). Thus, in the major periods of the English language we follow the pattern of the major classes of speech acts (for example as pointed out by Searle 1976). In the individual periods one should look at the major socio-

cultural specifications of the illocutionary acts and, of course, at their major manifestations.

Genres seem to require a descriptive, historical approach which looks at the prominent genres as reflected in the histories of the respective discourse domains and in the text documents which have come down to us. With genres, one needs to specify their functional and their linguistic profiles. In this way, the analysis could use a combination of a top-down and bottom-up analysis, as is suggested in the present paper. Thus, we approach speech acts (or text acts) through genres, embedding their different specifications and manifestation in the various genre-determined occasions of use; on the other hand the analysis provides a closer description of the linguistic profile of the genres and their functionally defined sections.

The series of synchronic descriptions of Old English, Middle English, Early Modern English and Late Modern English could later be complemented by tracing coherent developments and changes in genres and speech acts.

It goes without saying that we are still far away from such a comprehensive account. But our initial knowledge about speech acts and genres and the methodological considerations, for example, as shown in this paper, can serve as a fairly reliable starting point for such a project.

Notes

1 The number of studies on commissive speech acts in the history of English is more restricted, but see Arnovick (1999, 2006) and Pakkala-Weckström (2008).

2 On face and face-saving strategies see Brown and Levinson (1987).

3 On the exact composition of the underlying corpora in this and the following tables see Kohnen (2008).

4 The designations for the functionally specified sections are indicated in square brackets.

References

Archer, D. (2005), *Questions and Answers in the English Courtroom (1640-1760). A Sociopragmatic Analysis*. Pragmatics & Beyond New Series 135. Amsterdam: John Benjamins.

Arnovick, L.K. (1999), *Diachronic Pragmatics. Seven Case Studies in English Illocutionary Development*. Pragmatics & Beyond New Series 68. Amsterdam: John Benjamins.

Arnovick, L.K. (2006), *Written Reliquaries: The Resonance of Orality in Medieval English Texts*. Amsterdam: John Benjamins.

Bertuccelli Papi, M. (2000), 'Is a diachronic speech act theory possible?', *Journal of Historical Pragmatics*, 1(1): 56-66.

Biber, D. and E. Finegan (1992), 'The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries', in: M. Rissanen, O. Ihalainen, T. Nevalainen and I. Taavitsainen (eds.) *History of Englishes. New Methods and Interpretations in Historical Linguistics*. Berlin: Mouton de Gruyter. 688-704.

Biber, D., U. Connor and T.A. Upton (2007), *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure*. Studies in Corpus Linguistics 28. Amsterdam: John Benjamins.

Blum-Kulka, S., J. House and G. Kasper (eds.) (1989), *Cross-cultural Pragmatics: Requests and Apologies*, volume XXXI. Advances in Discourse Processes. Norwood, New Jersey: Ablex.

Brinton, L.J. (2001), 'Historical discourse analysis', in: D. Schiffrin, D. Tannen and H.E. Hamilton (eds.) *The Handbook of Discourse Analysis*. Oxford: Blackwell. 138-160.

Brown, P. and S.C. Levinson (1987), *Politeness. Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Claridge, C. (2000), 'Pamphlets and early newspapers: political interaction vs. news reporting', in: F. Ungerer (ed.) *English Media Texts – Past and Present. Language and Textual Structure*. Pragmatics & Beyond New Series 80. Amsterdam: John Benjamins. 25-43.

Claridge, C. (2001), 'Structuring text: discourse deixis in Early Modern English texts', *Journal of English Linguistics*, 29(1): 55-71.

Culpeper, J. and D. Archer (2008), 'Requests and directness in Early Modern English trial proceedings and play-texts, 1640-1760', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 45-84.

Culpeper, J. and M. Kytö (forthcoming), *Early Modern English Dialogues: Spoken Interaction as Writing.* Cambridge: Cambridge University Press.

Finegan, E. and D. Biber (1989), 'Historical drift in three English genres', in: T.J. Walsh (ed.) *Synchronic and Diachronic Approaches to Linguistic Variation and Change*. Washington D.C.: Georgetown University Press. 22-36.

Fries, U. (1994), 'Text deixis in Early Modern English', in: D. Kastovsky (ed.) *Studies in Early Modern English*. Berlin: Mouton de Gruyter. 111-128.

Fries, U. (1997), 'The vocabulary of ZEN: implications for the compilation of a corpus', in: R. Hickey, M. Kytö, I. Lancashire and M. Rissanen (eds.) *Tracing the Trail of Time. Proceedings from the Second Diachronic Corpora Workshop*. Amsterdam: Rodopi. 153-166.

Fries, U. (2001), 'Text classes in early English newspapers', *Early Modern English Text Types*. Special issue of *European Journal of English Studies*, 5(2): 167-180.

Fries, U. (2002), 'Letters in early English newspapers', in: K. Lenz and R. Möhlig (eds.) *Of Dyuersitie & Chaunge of Language. Essays Presented to Manfred Görlach on the Occasion of his 65th Birthday*. Anglistische Forschungen 308. Heidelberg: C. Winter. 276-289.

Jacobs, A. and A.H. Jucker (1995), 'The historical perspective in pragmatics', in: A.H. Jucker (ed.) *Historical Pragmatics. Pragmatic Developments in the History of English*. Pragmatics & Beyond New Series 35. Amsterdam: John Benjamins. 3-33.

Jucker, A.H. (ed.) (1995), *Historical Pragmatics. Pragmatic Developments in the History of English*. Pragmatics & Beyond New Series 35. Amsterdam: John Benjamins.

Jucker, A.H. (2000), 'English historical pragmatics: problems of data and methodology', in: G. di Martino and M. Lima (eds.) *English Diachronic Pragmatics*. Napoli: CUEN. 17-55.

Jucker, A.H. (2005), 'News discourse: mass media communication from the seventeenth to the twenty-first century', in: J. Skaffari, M. Peikola, R. Carroll, R. Hiltunen and B. Wårvik (eds.) *Opening Windows on Texts and Discourses of the Past.* Pragmatics & Beyond New Series 134. Amsterdam: John Benjamins. 7-21.

Jucker, A.H. (2008), 'Historical pragmatics', *Language and Linguistics Compass*, 2: 894-906.

Jucker, A.H. and I. Taavitsainen (2000), 'Diachronic speech act analysis: insults from flyting to flaming', *Journal of Historical Pragmatics*, 1(1): 67-95.

Jucker, A.H. and I. Taavitsainen (2008), 'Apologies in the history of English: routinized and lexicalized expressions of responsibility and regret', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 229-244.

Jucker, A.H., G. Schneider, I. Taavitsainen and B. Breustedt (2008), 'Fishing for compliments: precision and recall in corpus-linguistic compliment research', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 273-294.

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*. London: Longman.

Kohnen, T. (1997a), 'Toward a theoretical foundation of 'text type' in diachronic corpora: investigations with the Helsinki Corpus', in: R. Hickey, M. Kytö, I. Lancashire and M. Rissanen (eds.) *Tracing the Trail of Time. Proceedings from the Second Diachronic Corpora Workshop*. Amsterdam: Rodopi. 185-197.

Kohnen, T. (1997b), 'Text type evolution and diachronic corpora: historical writing in the history of English', in: M. Ljung (ed.) *Corpus-based Studies in English. Papers from the Seventeenth International Conference on Eng-*

lish Language Research on Computerized Corpora. Amsterdam: Rodopi. 153-166.

Kohnen, T. (2000a), 'Explicit performatives in Old English: a corpus-based study of directives', *Journal of Historical Pragmatics*, 1(2): 301-321.

Kohnen, T. (2000b), 'Corpora and speech acts: the study of performatives', in: C. Mair and M. Hundt (eds.) *Corpus Linguistics and Linguistic Theory: Proceedings of the 20th ICAME Conference*, Freiburg im Breisgau 1999. Amsterdam: Rodopi. 177-186.

Kohnen, T. (2001), 'On defining text types within historical linguistics: the case of petitions/statutes', *European Journal of English Studies*, 5(2): 197-203.

Kohnen, T. (2002), 'Towards a history of English directives', in: A. Fischer, G. Tottie and H.M. Lehmann (eds.) *Text Types and Corpora. Studies in Honour of Udo Fries*. Tübingen: Niemeyer. 165-175.

Kohnen, T. (2004), 'Methodological problems in corpus-based historical pragmatics. The case of English directives', in: K. Aijmer and B. Altenberg (eds.) *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* Göteborg 22-26 May 2002. Amsterdam: Rodopi. 237-247.

Kohnen, T. (2007), 'Text types and the methodology of diachronic speech act analysis', in: S.M. Fitzmaurice and I. Taavitsainen (eds.) *Methods in Historical Pragmatics*. Berlin: Mouton de Gruyter. 139-166.

Kohnen, T. (2008), 'Tracing directives through text and time. Towards a methodology of a corpus-based diachronic speech-act analysis', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 295-310.

Kryk-Kastovsky, B. (1995), 'Demonstratives in Early Modern English letters', in: A.H. Jucker (ed.) *Historical Pragmatics. Pragmatic Developments in the History of English*. Pragmatics & Beyond New Series 35. Amsterdam: John Benjamins. 329-344.

Maskell, W. (ed.) (1882), *Monumenta Ritualia Ecclesiae Anglicanae. The Occasional Offices of the Church of England*, vol. 3. Oxford: Clarendon.

McEnery, T., R. Xiao and Y. Tono (2006), *Corpus-based Language Studies. An Advanced Resource Book.* London: Routledge.

Nevala, M. (2003), 'Family first. Address and subscription formulae in English family correspondence from the fifteenth to the seventeenth century', in: I. Taavitsainen and A.H. Jucker (eds.) *Diachronic Perspectives on Address Term Systems*. Amsterdam: John Benjamins. 147-176.

Nevalainen, T. and H. Raumolin-Brunberg (eds.) (1996), *Sociolinguistics and Language History. Studies Based on the Corpus of Early English Correspondence.* Amsterdam: Rodopi.

Nevalainen, T. and S.-K. Tanskanen (eds.) (2004), *Letter Writing*. Special issue of *Journal of Historical Pragmatics*, 5(2).

Pakkala-Weckström, M. (2008), '"No botmeles bihestes": various ways of making binding promises in Middle English', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 133-164.

Schmied, J. and C. Claridge (1997), 'Classifying text- or genre-variation in the Lampeter Corpus of Early Modern English texts', in: R. Hickey, M. Kytö, I. Lancashire and M. Rissanen (eds.) *Tracing the Trail of Time. Proceedings from the Second Diachronic Corpora Workshop*. Amsterdam: Rodopi. 119-135.

Searle, J.R (1976), 'A classification of illocutionary acts', *Language in Society*, 5: 1-24.

Swales, J. (1981), *Aspects of Article Introductions*. Birmingham AL: University of Aston.

Swales, J. (1990), *Genre Analysis. English for Academic and Research Settings*. Cambridge: Cambridge University Press.

Taavitsainen, I. (1995), 'Interjections in Early Modern English: from imitation of spoken to conventions of written language', in: A.H. Jucker (ed.) *Historical Pragmatics. Pragmatic Developments in the History of English*. Pragmatics & Beyond New Series 35. Amsterdam: John Benjamins. 439-465.

Taavitsainen, I. (2005), 'Genres and the appropriation of science: loci communes in English in the late medieval and early modern period', in: J. Skaffari, M. Peikola, R. Carroll, R. Hiltunen and B. Wårvik (eds.) *Opening Windows on Texts and Discourses of the Past*. Pragmatics & Beyond New Series 134. Amsterdam: John Benjamins. 179-196.

Taavitsainen, I. and S.M. Fitzmaurice (2007), 'Historical pragmatics: what it is and how to do it', in: S.M. Fitzmaurice and I. Taavitsainen (eds.) *Methods in Historical Pragmatics*. Berlin: Mouton de Gruyter. 11-36.

Taavitsainen, I. and A.H. Jucker (2007), 'Speech acts and speech act verbs in the history of English', in: S.M. Fitzmaurice and I. Taavitsainen (eds.) *Methods in Historical Pragmatics*. Berlin: Mouton de Gruyter. 107-138.

Taavitsainen, I. and A.H. Jucker (2008), '"Methinks you seem more beautiful than ever": compliments and gender in the history of English', in: A.H. Jucker and I. Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 195-228.

Taavitsainen, I. and P. Pahta (eds.) (2004), *Medical and Scientific Writing in Late Medieval English*. Cambridge: Cambridge University Press.

Traugott, E.C. (2004), 'Historical pragmatics', in: L.R. Horn and G. Ward (eds.) *The Handbook of Pragmatics*. Oxford: Blackwell. 538-561.

Valkonen, P. (2008), 'Showing a little promise: identifying and retrieving explicit illocutionary acts from a corpus of written prose', in: A.H. Jucker and I.

Taavitsainen (eds.) *Speech Acts in the History of English*. Pragmatics & Beyond New Series 176. Amsterdam: John Benjamins. 247-272.
Werlich, E. (1976), *A Text Grammar of English*. Heidelberg: Quelle und Meyer.

The pragmatics of knowledge and meaning: Corpus linguistic approaches to changing thought-styles in early modern medical discourse

Irma Taavitsainen

University of Helsinki

Abstract

Pilot studies on the new specialized corpora with comprehensive materials, Middle English Medical Texts 1375-1500 *and* Early Modern English Medical Texts 1500-1700, *have already shown that the lines of development in the medical register are diversified, and a dynamic picture emerges. This study relates to the dissemination of knowledge and the negotiation of meaning across a wide selection of early modern medical texts. References to authorities with specific details are typical of the top genres of teaching and research in scholasticism and continue in the early modern period, but become adapted to new functions in more popular texts. In contrast, general references marking vagueness in medieval texts and occurring more frequently in texts for heterogeneous audiences acquire special meanings connected with the rising importance of discourse communities in the top genres of the seventeenth century. My approach is connected with historical pragmatics and historical discourse analysis. Corpus linguistic methods are applied to detect the overall trends and to locate relevant passages for qualitative analysis. For a more detailed microstudy, a keyword analysis is employed, with the frequencies of proper names as the prime point of interest.*

1. Discovering meanings of past texts and discourses

The broad view of pragmatics as a perspective on language use considers negotiation of meaning to be its core issue. Meanings are made at various levels, and pragmatic meanings depend on negotiation at the utterance level or even beyond, at the discourse and genre level, as will be demonstrated. Language external facts, including sociohistorical parameters, are important for interpretations. Language is viewed as an instrument that responds to, and is shaped by, communicative pressures in historical, ideological, social, and situational contexts (see Taavitsainen and Fitzmaurice 2007). Knowledge of the discourse community for whose benefit the text and the genre are created is of particular interest for studies on the diachronic development of professional languages.[1] Corpus linguistic studies, complemented by ideological, historical and textual assessments provide a firm ground to discover such meanings. My approach belongs to historical

pragmatics, an empirical branch of study which focuses on communication and contextual assessment of language use in past periods. New research questions and the desire to know more about past practices and cultural constructions are at the core of the approach.

2. Background, research questions and aim of the paper

In the history of science, the prevailing thought-style in the late medieval period was scholasticism.[2] Its defining feature was a firm reliance on axioms derived from ancient authorities, the source of knowledge was the quotative, 'that someone said so', and hearsay was the mode of knowing (see Taavitsainen and Pahta 1998). Authorities form the basis of scholastic knowledge, whereas empiricism relies on observation as the source of knowledge. The breakdown of the world view from medieval to early modern, from the Ptolemaic to the Copernican, marks the transition to a different approach to science, as the basis of knowledge changed through a broadening world and new environments. Authorities lost their status as the holders of truth and the reliance on their writings gave way to observation. The time between the mid-sixteenth and the mid-seventeenth centuries is crucial in this process.

What continued and what changed is an essential question in the history of science (Crombie 1994: 6). This statement serves as the point of departure in the analysis of scientific thought-styles in a diachronic perspective and the core issue can be approached by studying linguistic and stylistic change. This is a bold statement as, traditionally, linguistic analysis of scientific texts has been reduced to the role of finding antedatings such as earlier first occurrences of lexical borrowings.[3] This negative statement is provocative, and it is my purpose to show that it is possible to reveal more profound aspects of past cultures and changes in ways of thinking by linguistic analysis. My assessment focuses on how the old learned genres of scholasticism continue in the early modern period, and how their position changes in the course of time.

In this paper, my research questions deal with the larger structures of discourse, focusing on genres. How do old scholastic genres continue in early modern medical writing? Are they used in special ways? The hypothesis is that most late medieval genres and styles of writing continue with a gradual transition to the new mode of thinking. This is, however, an unknown area, and my intention is to trace some lines of development by corpus linguistic methods and find more detailed evidence of the mechanism of change. Genres guide the reception of texts and have a meaning-making function, as they are produced for the benefit and use of discourse communities. Genres constitute dynamic systems which undergo change and variation over the course of time as sociocultural needs change, and genres change accordingly: old genres are adapted to new functions, new genres are created, and genres that have lost their function cease to exist (Taavitsainen 2001a). An interesting point of comparison is provided by the circular movement detected as fluctuation and mutability in the stylistic features

of fiction in this period. The literature first targeted at aristocratic readers gained popularity with a common audience, e.g. John Lyly's style of writing became downgraded from fashionable and elegant style to old-fashioned and stilted writing and the same applies to romance (Taavitsainen 1993: 191; Margolies 1985: 14-16). This change in the stylistic value of discourse patterns may apply to other registers of writing and even to medical genre features, as this study shows.

The aim of this paper is to present some new insights and demonstrate a diversified and dynamic picture of early modern medical writing with the emphasis on the communicative function of genres in a sociohistorical frame.

3. *Early Modern English Medical Texts*, its text categories and genres of writing

The Scientific Thought-styles project group at the Research Unit of Variation, Contacts and Change at the University of Helsinki is compiling a register-specific corpus, *Early English Medical Writing 1375-1800*, to provide material for their corpus linguistic studies on changing thought-styles. Its second part, *Early Modern English Medical Texts 1500-1700* (EMEMT forthcoming), is used as data for this study. The first part, *Middle English Medical Texts 1375-1500* (MEMT 2005), is already available, and work on the third part, *Late Modern Medical Texts 1700-1800*, is also underway.[4]

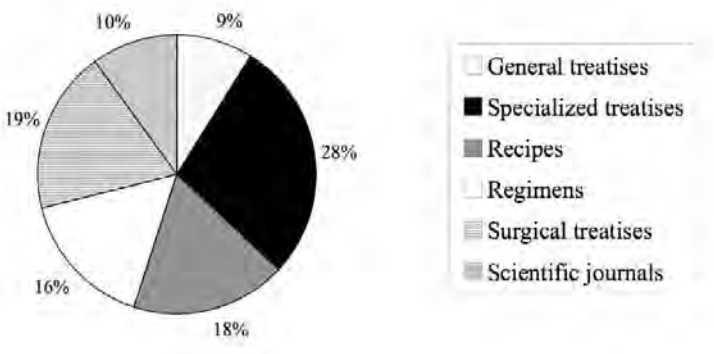

Figure 1: EMEMT categories

Our categories in MEMT (2005) are labelled 'surgical texts', 'specialized treatises', and 'remedies and materia medica'. This categorization was modified from the generally accepted tripartite division (see Voigts 1984). The first two categories belong to learned writing, the third contains texts belonging to the

remedybook tradition with more mixed origins (for more details, see MEMT: Introduction). For EMEMT, we surveyed the extant materials and the increasing variety of medical writing, and established our own categories according to the fields and topics of writing and types of publication so that it reflects medical history (see e.g. Siraisi 1990; Wear 2000).[5] The corpus is divided into the categories shown in figure 1 above, based on the fields and topics of texts.

Contrary to a common practice in corpus compilation, the above division does not take genres of writing as its point of departure.[6] Instead, the relation between the above categories and the genres of medical writing remains open and needs a great deal more work; the formation of genre conventions is one of the key research tasks of the scientific thought-styles research group.[7] Genres and their features in the early periods are very different from present-day genres of scientific and medical writing. According to my studies on MEMT, genre features were transferred from learned Latin writing in late medieval English. Latin genres provided a model for their vernacular counterparts, but when transferred, they lost their original institutional function and became modified, abbreviated and adapted in various ways, and in general the practical side became enhanced in the vernacular (Taavitsainen 2004). Doubtless the story continues in the early modern period as texts are built on others and intertextuality is an essential feature in early medical writing (see Mäkinen 2006).

4. Methodological challenges

I shall employ corpus linguistic methodology combined with qualitative contextual analysis. The new style of the *Philosophical Transactions*, especially the genre of research articles, received attention from its very start in the Royal Society period (e.g. Atkinson 1999; Gotti 2006; Valle 1999 and 2006; Mössner forthcoming), but other categories of medical texts in the same period have been neglected, perhaps because of the lack of readily available material. This is what EMEMT sets out to amend. The claim is that we can gain new accuracy with linguistic studies on new databases with carefully selected materials and an analytical grid which is specific enough to catch even minute differences. The methodological challenge for the present study is how to apply corpus linguistic methods in order to study genre dynamics and changing thought-styles. In an earlier study, systematic differences were detected between various layers of writing in references to authorities. We established a hierarchy of authorities in vernacular texts and correlated their names and the specificity of reference to the levels of writing. Learned and specialized texts referred to Galen and Hippocrates, Arab authors like Avicenna, Rhases, Haly Abbas and Averroes, and medieval Latin authors. Texts for heterogeneous audiences, by contrast, contained general references to doctors, physicians, authors, and leeches (Taavitsainen and Pahta 1998). In the present study on genre development in early modern medical writing, the empirical part builds on our previous study, but uses different, more modern, corpus linguistic tools in two methodologically separate studies.

In the first, the list of authorities made on the basis of the above-mentioned study gave an incentive to the present application with the KWIC (key word in context) concordance function of WordSmith4.[8] The problem of spelling variation was solved with the help of wordlists, which are easily created by the program. Various forms can easily and reliably be detected by screening the lists, e.g. HIPPOCRACTES, shows a great deal of spelling variation: *Hippocr*, *Hippocra*, *Hippocras*, *Hippocractes*, *Hippocrat*, *Hippocrate*, *Hippocrates*, *Hippocrate's*, *Hippocrates's*, *Hippocratus*, *Hypocras*, *Hyppocrates*, *Ipocras*, *Ipocrates*, *Ippocras*, and *Ypocras*. Other names also show variety, though not as much as the above example.[9]

The following authorities (i) and general categories of people (ii) were included in the present study, with their spelling variants:

i GALEN, HIPPOCRATES, AESCULAPIUS, AVICENNA, ALBUCASIS, RHAZES, HALY ABBAS, AVERROES, ARISTOTLE, PLATO AND PTOLEMY

ii PHILOSOPHERS, AUTHORS, PRACTITIONERS, PHYSICIANS, POETS.

To probe deeper into the special nature of references to authorities, a keyword analysis with WordSmith4 was performed for a measure of saliency of this feature. Keywords in this corpus linguistic application denote significantly more frequent words in a text (for a historical discourse application, see McEnery 2006). The focus was on category 2, 'specialized treatises', as the role of authorities would appear more clearly in a fairly strictly defined category. In general, three types of keywords occur. The most evident category is lexical items indicating the 'aboutness' of the text, words that reveal the topic. The second, perhaps more indicative of style and better at revealing the attitude of the author, for example, is the category of grammatical items. Proper nouns make up the third category. To my knowledge this group of keywords has not attracted much attention as it has been thought to be a "feature that perhaps isn't helpful […] although in some cases it may be interesting to pursue the use of proper names further" (Baker 2006: 207-208). The above-mentioned problem of spelling variation is more pronounced in the second empirical part of this study as non-standard spellings of words and frequent variation even in the most common lexical items prevent a straightforward application of the keyword function in historical material. The Variant Detector program, VARD 2, developed at the University of Lancaster to assist users of historical corpora in dealing with spelling variation, provides a solution to this problem as it is possible to normalize the spellings automatically.[10]

5. Genre dynamics

The top genres of scientific writing in the late medieval period consisted of commentaries and *compilationes*, encyclopaedic treatises, *questiones*, pedagogical dialogues, and *consilia* and *practica*, which reported case histories. University

teaching methods were based on these genres. The readership of learned vernacular texts consisted of people of high rank and professionals, such as highly educated physicians, surgeons, apothecaries and men of learning in general. The common people, 'the poor' and 'the unlearned', mentioned so often in the prefatory texts of early modern medical writing, especially in specialized texts and handbooks, was concerned with social decorum and did not reflect social reality as these texts hardly reached the illiterate (Taavitsainen 2002a). The most heterogeneous and widening readership was reached by almanacs, pamphlets, small booklets and writings of an encyclopaedic nature; yet according to extant external evidence, such books were also read by learned people and by the upper layers of society, that they reached the lowest social group of readers remains a matter of speculation. Literacy developments broadening the scope of potential readers and the aspirations of the rising middle classes are also relevant in this connection.

Commentaries and compilations were at the heart of the intellectual mainstream of scholasticism in both research and teaching. By definition, commentaries discuss opinions of ancient authorities pointing out differences and similarities in their writings, finishing with a summary and the author's own opinion, whereas compilations provided easy access to authoritative texts and disseminated knowledge to readers who could not access the originals (Minnis 1979). Compilations discuss opinions of ancient authorities, considering their differences and similarities, but the author's own opinion is not expressed. At the end of the medieval period, the top genres of scientific writing began to converge in Latin; in vernacular texts the genres overlap and merge from the beginning of the vernacularization period in the late fourteenth century (Taavitsainen 2004). The author's own opinion differentiated between the two top genres at first, but the distinction was already lost by the early modern period.[11] Furthermore, compilations had a bias towards instruction and practical knowledge and the technique of assembling quotations from authoritative texts merges with book-making practices in commonplace books, i.e. private collections of useful texts and extracts were compiled in notebooks for people's own interest, amusement and instruction (Ayoub 1994: 5-6; Taavitsainen 2005: 187 and forthcoming a). In general, textual histories are extremely complicated as texts of any genre could be subjected to commentary, and hybrid texts combining textual forms can be encountered. An illustrative example is found in a vernacular text from the end of the sixteenth century (example (1)). It is a recipe following the conventions of the genre in part (1a), with the imperative *take* in the opening position giving a strong genre signal, followed by the ingredients listed with measures. The text is, however, a hybrid form, as part (1b) fulfils the criteria of a fully-fledged commentary, with quotations from authorities and the author's own opinion stated at the end. Part (1c) states where the medicine is available and is a forerunner of the new emerging genre of medical advertisements. This example shows how complicated the genre map can be and how our corpus categories and genre labels concerning early scientific writing do not coincide.

(1a) Take of the three kinds of pepper +Q .6. +q .ij. ginger, anise seedes, thyme ana. +Q .ij. +q .ij. spikenard, amomum ana. +Q .j. +q .j. cassia lignea, asarum, enula camp. dried, persley seeds seseli ana. +q .ij. saccharum albissimum q. s.

(1b) This is the same laborious confection set downe by Mesue, of which Galen maketh mention, lib. 4. cap. 11. De sanit. tuend. For Galen in that place doth recite by name euery simple: saue that Galen in stead of Amomum, mentioneth the seeds of ammios. And where our authors do differ in opinion, what is the right Amomum, and all writers do agree, that we haue the right ammi: I adiudge it better in these daies to put in this receit the seeds of ammi, according as Galen prescribeth, than amomum, as Mesue counselleth.

(1c) So you haue both medicines in vse in our time made of three peppers, and (f._B8) are to be sold in the Apothecaries shops, vnder the name of Diatrion pipereon. (Walter Bailey, *Three Kinds of Pepper*, 1588, f.B7v -8r)

The above tendencies, hybrid forms and applications of learned models to texts of various types can be verified by qualitative studies. The above example is from a learned text, and it can be assumed that more examples can be found in various categories of the corpus. These observations provide the point of departure for my corpus linguistic studies in search of more evidence on the circular movement of stylistic features from top genres of scholasticism to supposedly more popular and less technical writing. The dynamics of genres with their circular movements are depicted in figure 2.

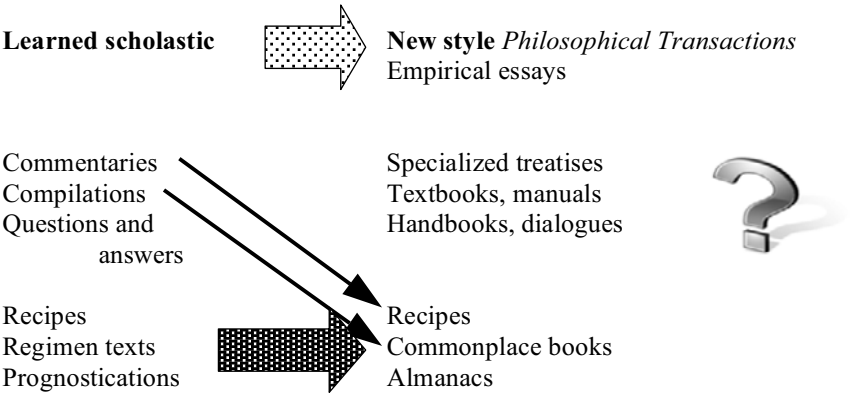

Figure 2: Genre developments from late medieval to early modern medical writing

6. References to authorities according to text category and genre of writing in EMEMT

A clear chronological development can be verified in the overall charts (figures 3 and 4). For a more analytical view, an assessment was made according to the textual categories of the corpus. The overall trends confirm the transfer from scholasticism to empiricism, from reliance on authorities to the rising importance of the discourse community in creating new knowledge through observation and communicating the results of experiments within closed circles of scientists.[12]

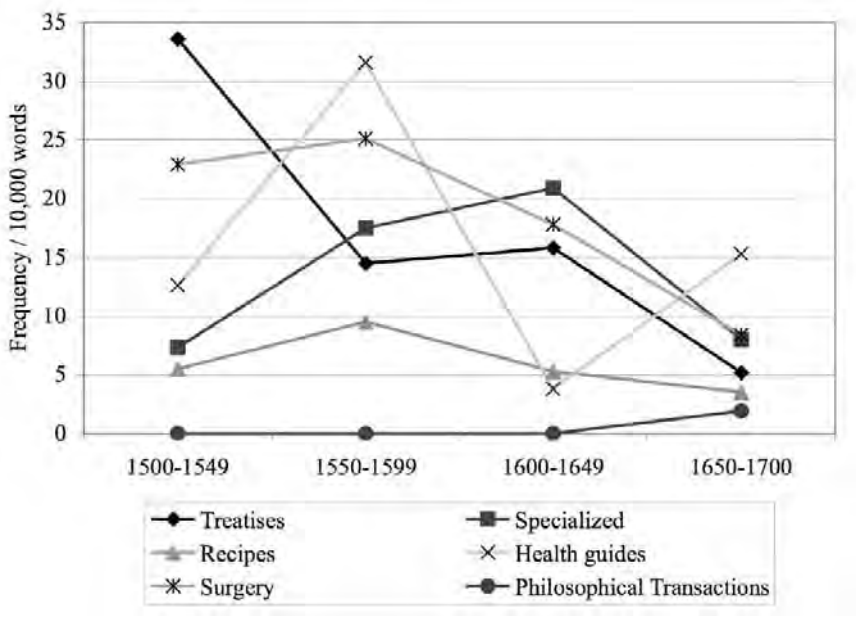

Figure 3: Frequencies of references to authorities according to textual category

These general tendencies (figure 3) are indicative of the decline in the reliance on authorities, but a more analytical assessment according to textual categories reveals a more complicated pattern.

KWIC concordances of proper names showed that authorities occur in the quotative mode with speech-act verbs of saying. More interestingly for the present purpose, the concordances show the distributions of references to authorities at the text level. Texts with frequent quotations and clusters of authorities indicate passages potentially in accordance with genre characteristics of commentaries in the medieval tradition, but qualitative assessment is needed at this stage. Locating such passages with computer-aided methods proved useful for identifying passages written as commentaries in the scholastic style in

EMEMT. There are several lengthy passages in accordance with the medieval tradition, and although the texts are not labelled 'commentaries', they can nevertheless be considered the aftermath of the medieval top genre of research and teaching. Texts for professional audiences in categories 1, 2 and 3 all contain commentaries in the old style. One of the texts in category 2 also shows adaptation to the more popular direction. Category 4 proved different, as the commentary style is used for new stylistic purposes in some texts, and the scholastic pattern of argumentation with references to authorities is developed further in a seventeenth-century handbook (see example (4) below). There is some variation in category 5, as references abound in teaching dialogues and textbooks that contain discussions on the teaching of ancient authorities. In general, there is more variation in style in this category, but the core issue is the same and the dialogue frame is only fairly loosely imposed on the commentary form, as shown in example (8) below. The categories will be assessed in order and illustrative examples of text passages quoted to demonstrate commentary features in texts.

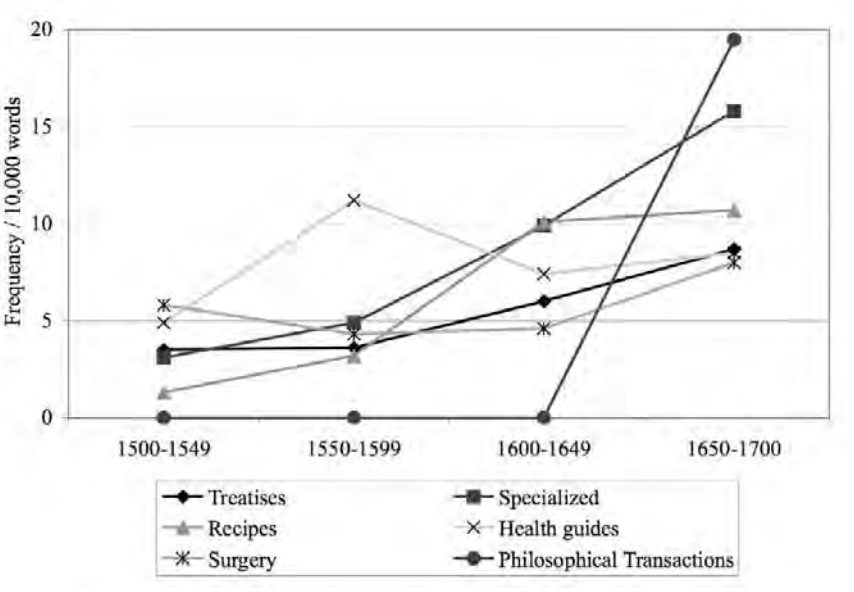

Figure 4: Frequencies of general titles according to textual category

6.1 Comprehensive medical treatises (textbooks)

Texts in this category include overall surveys or systematic accounts of medicine as a whole, with 16 texts and some 146,000 words in all. The coverage is fairly uneven across the fifty-year slots (see figure 5).

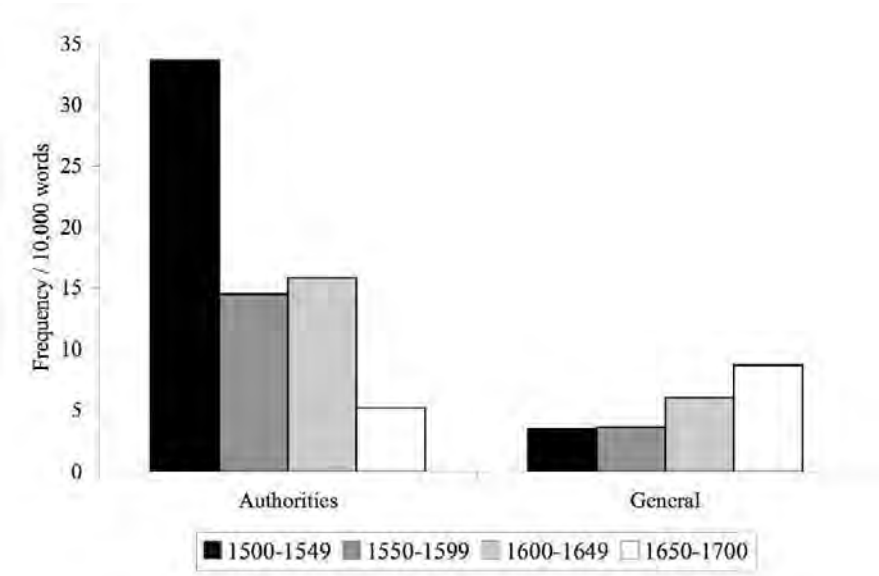

Figure 5: Frequencies according to textual categories: category 1, general treatises or textbooks

The first half of the sixteenth century shows a very high frequency, but after that the tendency declines rapidly. The peak coincides with the beginning of the period and shows that the old thought-style continued.[13] My more recent study focusing on the scale from learned to 'popular' (Taavitsainen forthcoming a),[14] showed that this category contains heterogeneous texts as the target audiences vary: there are texts for professionals, texts for learners of the profession and texts for domestic use targeted at women. They fall at very different places on the scale from learned to popular. General references in the medieval style remain at a low level all through the assessed period, with a slight increase towards the end. These references are in accordance with the rising importance of the discourse community, and few general references with vague referents were found.[15]

6.2 Specialized treatises

Texts dealing with specific fields of medicine, substances or methods of treatment belong to this category. It covers more than a quarter of the corpus, with 59 texts and some 494,000 words. The pattern is very different from category 1 as the first part of the sixteenth century produced few such texts; e.g., Caius' *The Sweating Sickness* (1552) is the first specialized treatise focusing on one disease.

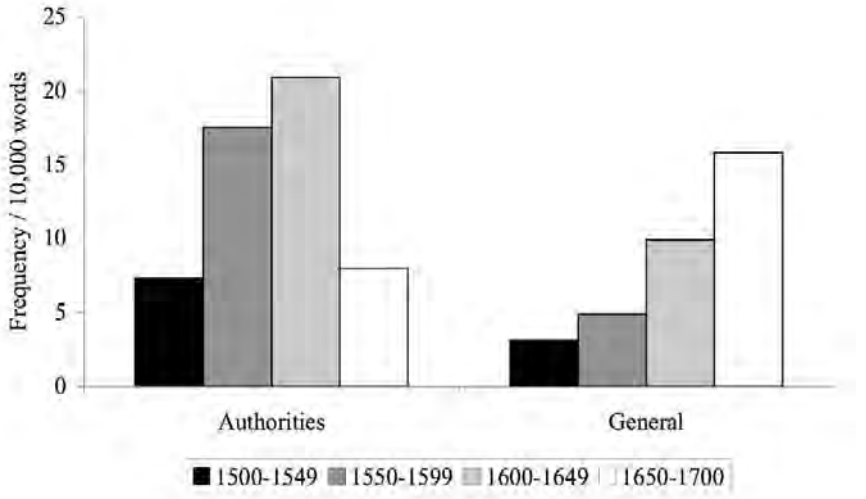

Figure 6: Frequencies according to textual categories: category 2, treatises on specific topics

References to authorities become more frequent in the latter half of the sixteenth and the first half of the seventeenth centuries, but there is a sharp decline after this. In figure 6 above, all texts in the second category are treated together, but there are differences between the subcategories. Illustrative examples from categories 2b and 2d are quoted below. Clusters of references proved reliable indicators of passages with the old commentary style. The passages quoted below are in accordance with genre features of top scholastic science: deontic modality prevails, passive voice sentences are frequent, there is code-switching to Latin, and the references are specific.

The following text of subcategory 2b contains frequent references to authorities. The style of writing resembles that found in the Middle English *Phlebotomy* from the first phase of vernacularization of academic texts. Deontic modality prevails in both texts (italics added in all examples).[16]

(2) But if our scope and purpose be only simply to vent, then *it is best to do* it by letting a little bloud, and often according to the rule of Auicenna, Melior est multiplicatio numeri quam quantitatis. Otherwise if the case be compound that both there be a fulnesse and a boyling, that we *must* both euacuate and vent, *then shall it be most fit to do it at once*, and plentifully and as long as the strength will permit, as is taught at large by Galen in the eight booke of his Methodus medendi. And in the same booke speaking of some agues that are like to Diarian feuers, and do come of obstructions, he doth vse these words: That the humour may be vented *wee haue neede of the great remedy, wee must let bloud* the party being of sufficient strength,

> although there be no signes of plenitude. How it shal be knowen when the humours do thus boyle and haue neede of venting, it shall be layd open at large in the two Chapters next following. (Simon Harward, *Harwards Phlebotomy: Or, A Treatise of letting of Bloud*, 1601, pp. 4-5)

The following passage from the year 1612 contains a somewhat startling narrative embedded in the text and attributed to an authority. The textual form is in agreement with the medieval commentary style in which quotations from authorities could be argumentative, narrative or expository. The majority of references are specific. This text belongs to subcategory 2d:

(3) *Aristotle in his booke De generatione Animalium*, is of opinion, That bruit beastes going with young, are not subiect to any diseases: and contrariwise, that Women are verie often sicke. *Hippocrates saith*, That they be pale and wan, to shew that they are subiect to many infirmities. In times past when men and women were sold like slaues, if there were any found that were with child, she was not warranted for whole and sound by him that sold her, as *Vitruuius writes* in his second book; because they were troubled and subiect to so many diseases. (Iames Gvillemeav, *Child-birth or, the happy deliverie of vvomen*, 1612, pp. 32-33)

The following example comes from a commentary passage in Aristotle's *Masterpiece* (1684), perhaps the most popular and widespread encyclopaedic handbook of reproduction and sexual matters in the seventeenth century.[17] The passage provides further evidence of the development of commentary features for new purposes. The argument concerns a moral issue:

(4) 'tis possible, and has been frequently known, that Children have been born at 7 months; but the matter being wholly left by the Lawyers, who decide Controversies to the Physicians to judge […] whether […] 7, 8, 9, or 10 months. *Paul the Counsellor has this Passage in his nineteenth Book of Pleadings*, viz. It is now a received truth, that a perfect Child may be born in the 7 month, *by the Authority of the learned Hypocrates*. And therefore we must believe, that a Child born at the end of the 7th month, in lawful Matrimony, may be lawfully begotten. *Gallen, in the 6th Chapter of his third Book*, handleth this Argument, but rather according to Mens Opinions than according to the truth of the business, or from natural Reasons, who supposeth there is no certain time set for bearing Children. And *from the Authority of Pliny* […] *saith Lemnius*. I know many married People in Holland... (*Aristotle's Masterpiece*, 1684, p. 70)

The reference in the old scholastic style is employed to convince and persuade ("It is now a received truth […] by the Authority of the learned Hypocrates. And therefore we must believe…"), and to support the claim that full-time newborns were possible within a shorter time than the normal nine months. This reflects

societal norms and shows an attempt to provide an honourable way out of moral judgements.

General references to 'philosophers', 'wise men' and other such groups in category 2 show the opposite tendency to that of giving references to authorities in the latter half of the seventeenth century. An example of the frequent use of general titles is taken from a text written in 1699, the very end of the corpus period (example (5)). It is a passage from a contemporary controversy between men of learning, referred to with the general title of "physicians".[18] It can be assumed, however, that the readership knew exactly who the people concerned were, and thus these references are very different from their medieval counterparts.

(5) THO I have been much solicited, to shew my Opinion, about the Debate *betwixt the two Physicians* […] and violently oppos'd by a certain Club of Physicians; I yet delay'd to give my Sentiments therein, until I should see whether the Learn'd Colledge of Physicians would interpose therein, […] he was happily treated that way, *by the Joint Advice of the Physicians* who waited on him, and that at that time *few Physicians* approved of […] Dr Sydenham in his last work… (Andrew Brown, *The Epilogue*, 1699)

6.3 Herbals, antidotaries and dispensatories

Texts of this category include instructions for preparing medicines and various other substances, collections of recipes, and herbals. It comprises 38 texts and some 308,000 words.

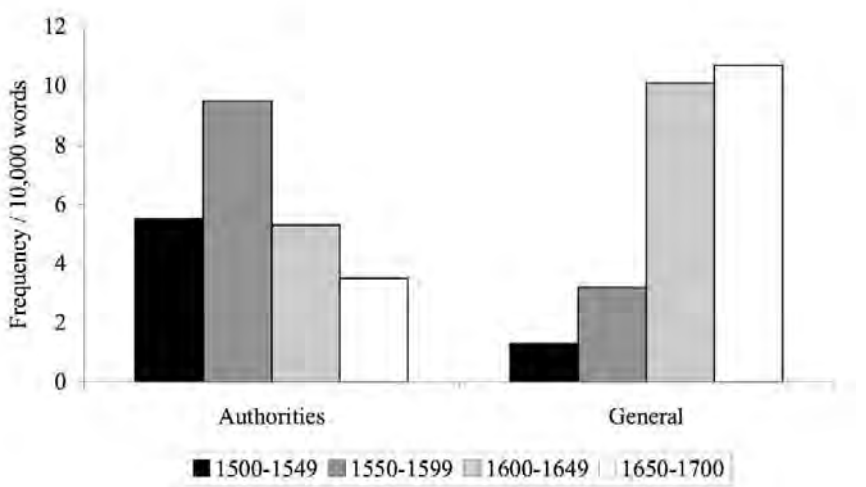

Figure 7: Frequencies according to textual categories: category 3, recipe collections

The overall pattern of frequencies is different in this category (see figure 7 above). In the first half of both centuries frequencies are at approximately the same level, there is a peak of references during the latter half of the sixteenth century, and a decrease follows towards the end of the assessed period. General references in this category show a prominently rising tendency with low frequencies in the sixteenth century and an abrupt rise in the seventeenth century.

6.4 Regimens and health guides

This category comprises texts focusing on the good life and the preservation of health and they provide advice on diet (food and drink) and cover various aspects of lifestyle, such as exercise, sleep and sexuality, or discuss what constitutes a healthy environment. Texts of this category are 29 in number, with some 281,000 words, mostly intended for heterogeneous audiences (see figure 8).

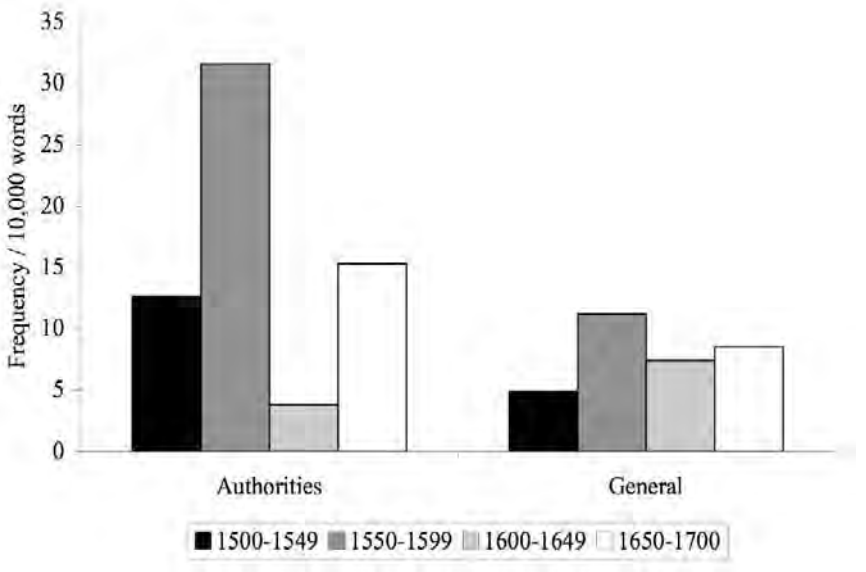

Figure 8: Frequencies according to textual categories: category 4, regimens and health guides

References to authorities seem to have gained additional functions in this category: commentary style is employed to lend an aura of learnedness to the contents, and this stylistic function overrides the informative function of indicating the source of knowledge. The year of composition of the following text (example (6)) is 1558-1559. According to my observations, this passage shows the earliest modification of scholastic style to a new function. Besides giving information, the references create an image of Humphrey in accordance with the

tradition of wisdom literature as a learned teacher, fit to instruct a foolish young man in matters of health (see Taavitsainen forthcoming c).

(6) *Hum.* Euery thing hath his time conuenient, and must bee done with sober discretion, and not with rash ignorance, […] Therefore the cause must be knowne, and the time obserued, *as Galen writeth in the Commentarie of the Afforismes of Hippocrates*, manie bodies be extinguished by sudde~ death, in whom is extreme fulnesse, or aboundance. For aboundance of blood or any other humor *sayth Aristotle*, is the cause of many sickenesses, […] *as Galen sayeth*. The letting of bloud dryeth vppe the superfluous moisture […] Vnto this *agreeth Rasis*, saying, it helpeth greatly against Leprosie, Squinances, Appoplexes, Pestilences, &c. […] as *Rasis sayth*, the spring of the yeare is the chiefe time to let bloud […] For *Hippocrates sayeth*, without doubt it is needfull to purge the superfluities of the bodie. (William Bullein, *The Gouernment of Health*, 1595, ff. 19v-20r)

General references in this category remain low all through the two hundred-year period, but some interesting examples can be found. The following passage gives evidence of the growing importance of the discourse community and shows a different epistemic stance even before the Royal Society period (see Taavitsainen 2001b). Francis Bacon wrote in 1638:

(7) *PHilosophers might better than Physitians* follow common opinion in condemning many Services and Messes of meate, lengthning not Life, but preserving health, for a Heterogeneous mixture of meates doth more readily nourish the veines, breeding better moysture than one kinde of meate: moreover, variety excites the Appetite, and the Appetite sharpens Disgestion. (Francis Bacon, *The Historie of Life and Death*, 1638)

6.5 Surgery and anatomy

This category contains descriptions of human anatomy and books focusing on various aspects of surgery, with 32 texts and some 318,000 words. The works in this category are mostly textbooks as they are targeted at learners of the profession. The dialogue form became common in the latter half of the sixteenth century. In example (8) below, the names of the discussants are taken from real life.[19]

General references remain at a low level throughout the period in this category with little difference between the fifty-year slots (figure 9).

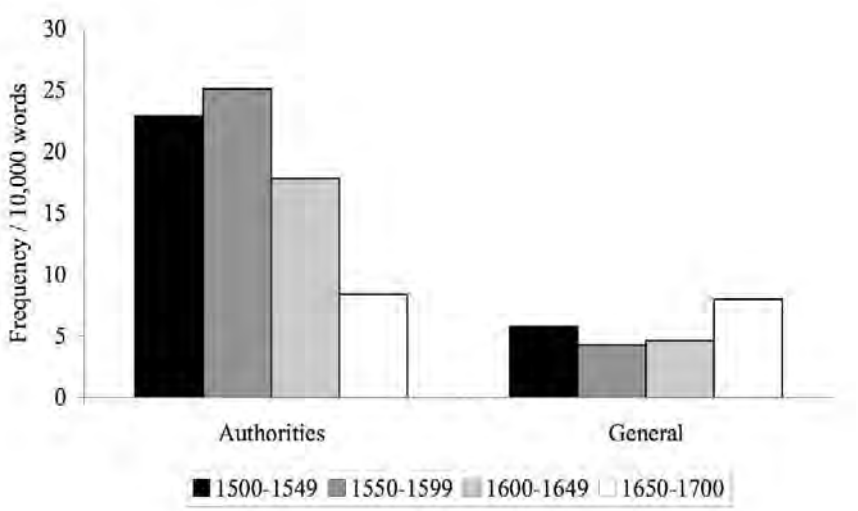

Figure 9: Frequencies according to textual categories: category 5, surgical and anatomical treatises

(8) *Iohn Feilde.* What parte of Hippocrates, Galene, Auicenne, Paulus, Rhasis, Albucasis, and Haliabbas, haue you rede? theis be of greter authoritie, and of whom you shal learne sounde doctrine.
Iohn Yates. Wyth theis I haue not ben much acquaynted.
Tho. Gale. Well, then the nexte waye is that you do put oute suche questions as you wolde haue answered, and stande in doute: and also answere to that which shalbe demaunded.
Iohn Yates. Right gladly, for ther is no waye by whyche I shall so muche, and in so shorte a tyme profit.
Iohn Feilde. Truth it is, wherfore let vs begin wyth out further detractynge of tyme. (Thomas Gale, *An Institution of a Chirurgian*, 1563, f. 2)

6.6 Journals

This category is defined by the medium and is very different from the other categories. Its texts, 139 in all, covering some 164,000 words, are selected from the first scientific journal, the *Philosophical Transactions* of the Royal Society. These texts are mostly short articles, reports or reviews. They are connected with medical topics, including physiology and anatomy, the nature of medical substances, accounts of medical experiments and discoveries, and reviews of medical books.

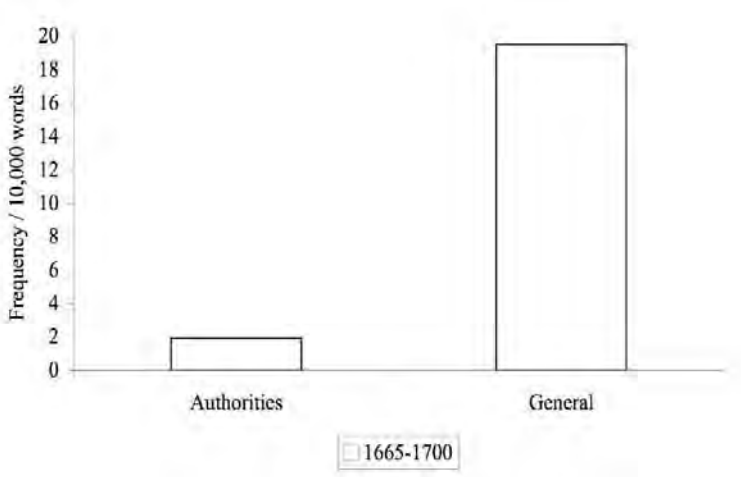

Figure 10: Frequencies according to textual categories: category 6, scientific journals

Figure 10 is different from the rest and should be contrasted with the last columns of the other figures, as the period covers only the final decades 1665-1700. The results obtained in this category are as might be expected: authorities are low in frequency and the discourse community with general references is prominently present, as the following example shows:

(9) Decemb. 19. we used what art we could to dispose the fancy of our Patient to suffer the Transfusion, which we resolv'd should be tried upon him that night about 6 a Clock. Many persons of Quality were present, together with *several Physicians and Chirurgeons*, too intelligent to suspect them of being capable of the least surprise. M. Emmerez open'd the Crural Artery of a Calf, and did all the necessary preparations in their presence; (*Philosophical Transactions*, 1667-1668_pt2_page 620)

The function of general references in the above passage, for example, is the opposite of the medieval "physicians" with vague referents and generalized meanings even in top level treatises, e.g. a fourteenth-century passage reads: "And what is ane arterie? – no þing elles but a place of spirituel blode; and þat is knowen to alle phisiciens & surgenes" (MEMT Chauliac, *Anatomy* (interpolated version), p. 14).

7. Triangulation with corpus-driven analysis: keywords (WordSmith)

In the above survey, category 2 contained several texts with clusters of references to authorities, which in qualitative assessment exhibited medieval commentary features. In order to discover more about this category, I decided to scrutinise a selection of texts more closely by applying keyword analysis (see above). My hypothesis was that some special features of the mechanism of change could be detected when assessed within a narrow scope in one category with a wide coverage, and with subcategories according to the topic of writing. A herbal text from 1588, *Three Kinds of Pepper*, serves as a case study and an illustration of the method,[20] and it is the same text that contains a hybrid passage of a recipe with a commentary quoted in example (1). The treatise was written by Walter Bailey (1529-1593), a highly learned doctor of medicine and Fellow of the College of Physicians. The text discusses the pepper tree and points out how eye-witness accounts differ from the received truth of ancient authorities. The plot view indicates the discourse positions of the references and thus it provides more detailed evidence of the mechanism of change in scientific thought-styles as an early advocate of the changing premises of scientific knowledge (see figure 11).

The keyword list (WordSmith4) of *Three Kinds of Pepper* (1588) is in accordance with the general expectations (see above). The words at the top of the list, those significantly more frequent than found in the reference corpus, were about the topic and included: *pepper*, *pipereon*, *peppers*, *ripe*, *black*, *medicine*, *tree*, *white*. This list reflects the earlier conceptions that black and white pepper represented different degrees of ripeness of one plant. The second category of grammatical items was not very helpful in this case (*do*, *doth*), but the third category of proper nouns proved interesting and worth pursuing further (cf. Baker 2006: 207-208). Galen tops the list of authorities with 52 occurrences. An abbreviated reference to his work *De sanit. tuend* is also near the top of the list, and such precise references can be taken as an indication of the learnedness of the text. Next in frequency is Dioscorides (c.40-c.90 BC) with 18 hits. He was a Greek physician, a pharmacologist and surgeon, and the father of herbal knowledge. Third is Garcia with eight occurrences. The name refers to Garcia de Orta (1501/2-1568), who had sailed for India in 1534 and become known for his expertise in spices, medicine and tropical diseases. The list continues with Nicolaus Myrepsicus (six), Serapio (five), Auicenna (five), Monardes is referred to twice, and Auerrhois, Clusius, Hippocrates and Matthiolus occur once. Most authorities are derived from classical antiquity, and it is not surprising to find Dioscorides in a prominent position. In contrast, Garcia and Monardes are contemporary authors. The point of interest is the dichotomy between the old authorities, representing received knowledge from scholasticism, and the advocates of new knowledge by observation. The positions of the proper names are interesting: Galen is most frequent towards the end, Dioscorides and Gracia at the beginning. Of particular interest are the locations where the new and the old intersect. The new is presented first, with emphasis on the innovation.

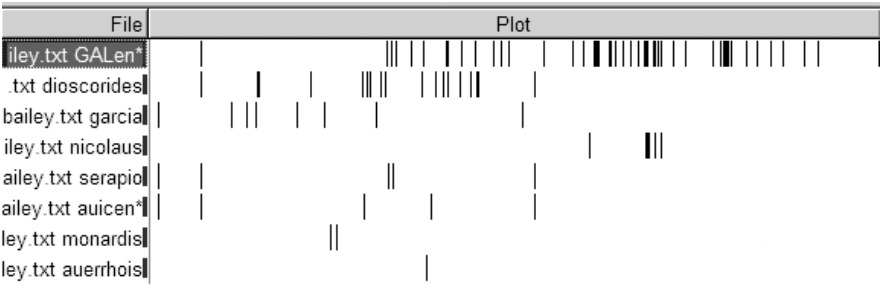

Figure 11: Authorities of *Three Kinds of Pepper* (1588): plot view

A qualitative assessment of the crucial passages confirms the polemical contrast focusing on the source of knowledge, whether observation or book learning, and the growing disbelief is openly discussed. The argumentation is powerful with parallel structures and affective patterns ("how greatly", "it is evident and well knowen").

(10) The nauigations in these latter yeeres made by the Portingales into the east Indians, and by the Spaniards into the west Indians, hath made manifest to vs, *how greatly the old authors*, I meane Dioscordes, Galen, Plinie, Auicenna, Serapio, and other writers of the former time *were deceived* […]. For first, all the ancient writers in their monuments haue deliuered to vs […] But *by the nauigations* of the Portingals, and of the Spaniards into those countries, in which these pepper trees do growe, *it is euident and well knowen*, that […] And Garcia, who lived in these parts, writeth… (Bailey, *Three Kinds of Pepper*, 1588, f. A5r)

The mechanisms of change can be assessed with a magnifying glass with the key word function, using the plot view, but qualitative assessment is necessary to complement and interpret the statistical evidence given by the program.

8. Conclusions

The basis of medicine changed with increasing knowledge of anatomy and physiology. The patterns of authorities versus general references, as well as the names of contemporary and ancient authorities, reflect changing thought-styles. The medieval usage of general titles showed vagueness, but this function is not prevalent in the early modern period. In the seventeenth century, the locus of scientific knowledge changes as the discourse community and its members become important. General references gain new functions: they become specific in the sense that all members of the discourse community knew who they referred to, enhancing the solidarity by the shared common ground, as is well demon-

strated in example (5). In the material presented here, features of the new style of writing science can be found several decades before the first scientific journal was founded. It shows that the new style of writing is not only a feature of the *Philosophical Transactions*, but already occurs more broadly in earlier medical writing.

In literature about the scientific revolution of the seventeenth century, it has been pointed out that there was no single coherent cultural entity called 'science' in the seventeenth century that underwent revolutionary change, as different fields of science proceed at different paces. Instead, science in the early modern period can be described as a diverse array of cultural practices aimed at understanding, explaining and controlling the natural world, each with different characteristics and each experiencing different modes of change (Shapin 1996: 3, 68-69). The present study shows this statement to be true at the discourse level. Old patterns continue in the first half of the sixteenth century without much change, as the old commentary style lives on beyond the period of its heyday. In most text categories there seems to be little change and, for example, in the textbook category the style remains fairly stable throughout the assessed period. This is in accordance with how contemporary authors saw themselves and their work in terms different from the modern perspective, for example Andreas Vesalius (1514-1564), celebrated as the inventor of rigorous observational methods, saw himself as reviving Galen's knowledge (Shapin 1996). Changes seem to have been initiated in the latter half of the sixteenth century with the broadening horizons of the world, new knowledge based on observation and a growing lack of faith in the wisdom of the ancient authorities. The broadening of the world created a new trend in the literature with eye-witness accounts giving different evidence about nature than the descriptions in inherited accounts. This discrepancy is openly discussed in the case study from 1588. New knowledge diffused slowly through layers of scientific writing, working within the old frame, but combining with the new modes of thought.

The circular movement attested in early modern fiction is also found in scientific writing, though the stylistic features are very different.[21] The early modern period provides crucial material for studying changing thought-styles: old practices continue as such until the seventeenth century, with some variation in the style of writing; but at the same time the old scholastic way of writing gains new stylistic functions. It is used to lend an aura of learning to texts for more heterogeneous and popular audiences, and this use continues for centuries as, for example, *Aristotle's Masterpiece* continued to be printed until the twentieth century. The new style in medical and scientific writing in pamphlets and other treatises for a wider audience is an under-studied area, and more evidence is needed to construct a clearer view of what took place and how, and to what extent, thought-styles changed in early modern medical discourse.

Notes

1 Swales's (1990: 29) original concept of discourse communities with six defining criteria has been modified to correspond with, and reflect, late medieval practices as "groups of people connected by texts" (Jones 2004: 24). More information is extant from the early modern period and the importance of discourse communities increased over the course of time, as the present assessment shows.

2 Thought-styles can be defined as underlying commitments in ways of thinking and making decisions (Crombie 1994: 6).

3 Grmek (1998: 18) writes: "In the field of philology, the task is certainly fascinating, but there are no major methodological innovations, and the work fits into the tradition of the fight against anachronisms".

4 The current project members are: Irma Taavitsainen and Päivi Pahta (project leaders); Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr, Jukka Tyrkkö, Alpo Honkapohja (doctoral students); Anu Lehto (research assistant); Martti Mäkinen (post doctoral researcher, Stavanger). For further information, see www.helsinki.fi/varieng/CoRD/corpora/CEEM/index.html.

5 We studied the literature and consulted experts in the history of medicine. We are grateful to Dr Peter Jones of King's College, Cambridge, for his advice over the years.

6 For example, the *Helsinki Corpus* and ARCHER are multigenre and multiregister corpora and structured according to the genre principle, see http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html and Biber et al. (1994).

7 See http://www.helsinki.fi/varieng/domains/scientific%20thought.html.

8 WordSmith5 became available only after this study was completed. The search word is the key word in this application.

9 Upper case is used to indicate that the forms in the list include all variant spellings, e.g. HIPPOCRATES includes the variants marked with lower case italics, above.

10 VARD 2 is an interactive software program using techniques derived from modern spell checkers to replace spelling variants within historical texts with their modern counterparts. Category 2 of EMEMT was processed with VARD 2 for the present purpose. I am grateful to Paul Rayson and Alistair Baron for their help in this process. See http://www.comp.lancs.ac.uk/~barona /vard2/.

11 For simplicity, I use the more technical term 'commentary' for this new merge.

12 Discourse communities were more heterogeneous and less restricted to specialized fields in the early modern period than present-day scientific communities. In the Royal Society, all members took part in medical experiments and discussions, as the *Philosophical Transactions* testify.

13 Period divisions in language histories are always arbitrary. In our corpus compilation, we accepted 1500 as it is generally considered the dividing line between the late medieval and early modern period. In medical writing, 1550 would have been more accurate in many respects, as in the first half of the sixteenth century medieval traditions continue and new trends emerge only in the middle of the century.

14 The study deals with commonplaces of science and humoral theory. Applications prevail at the more popular end, e.g. in the handbook targeted at women and household use, but in the present research task the differences between the texts of category 1 were less striking.

15 For example, in phrases like: "Physicians call this an imperfect Palsey…" (Bruel, *Praxis Medicinæ*, 1632, p. 13).

16 For comparison, the following passage from the end of the fourteenth century is given here: "Yt is to wyt as seyþ Galien in Metategni þat if it owth to be done by apoferisim, be þe flebotom dippyd in oyle þat the wond of þe flebotomie be leng holdyn oppyn. Also sum men of custum when þey are flebotomyd swoneþ; sech, as seþ Constantyne, be þey refreshyd..." (MEMT, *Phlebotomy*, early fifteenth century p. 39).

17 This text illustrates how heterogeneous our corpus categories can be: it is a specialized text of category 2d as it deals with sexual matters and reproduction, but at the same time it represents the popular end of the scale in early modern medical writing.

18 Controversies are typical of early modern scientific and religious writing (see e.g. Fritz 2008). In the field of medicine, several texts in this category focus on polemical issues (see Wear 2000). The tobacco controversy, for example, was equally concerned with economic motivations (Ratia, forthcoming).

19 Fictional dialogues occur as well, e.g. in Bullein's *Governayle* quoted above in category 4, the discussants are Humphrey and John. *The Fever Pestilence* by the same author grows into a social satire with Latinate names according to the medieval tradition. This text is included in the appendix "Medicine in Society" in EMEMT.

20 As a preliminary step, the spellings of the texts in this category were normalized (see above), and the analysis was performed with the rest of the category as the reference corpus. A qualitative reading had already revealed that this text contains important evidence of changing thought-styles (Taavitsainen 2002b) and how news of the new world was disseminated (Taavitsainen forthcoming b), but the plot view adds to our knowledge in showing the pattern of argumentation in more detail.

21 Perhaps the tendency for writing styles to become downgraded over the course of time is universal, but further studies are needed for more definite conclusions.

Corpora

Early Modern English Medical Texts 1500-1700 (EMEMT forthcoming). Compiled by Irma Taavitsainen, Päivi Pahta, Turo Hiltunen, Ville Marttila, Martti Mäkinen, Maura Ratia, Carla Suhr and Jukka Tyrkkö.

Middle English Medical Texts 1375-1500 (MEMT 2005). Compiled by Irma Taavitsainen, Päivi Pahta and Martti Mäkinen. CD-ROM. Amsterdam and Philadelphia: John Benjamins.

References

Atkinson, D. (1999), *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Mahvah, New Jersey: Lawrence Erlbaum.

Ayoub, L.J. (1994), *John Crophill's Books: An Edition of British Library MS Harley 1735*, Unpublished PhD thesis, University of Toronto.

Baker, P. (2006), *Using Corpora in Discourse Analysis.* London and New York: Continuum.

Biber, D., E. Finegan and D. Atkinson (1994), 'ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers', in: U. Fries, G. Tottie and P. Schneider (eds.) *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993*. Language and Computers: Studies in Practical Linguistics 13. Amsterdam and Atlanta, Georgia: Rodopi. 1-13.

Crombie, A.C. (1994), *Styles of Scientific Thinking in the European Tradition*, 3 volumes. London: Duckworth.

Fritz, G. (2008), 'Communication principles for controversies – an historical perspective', in: F. van Eemeren and B. Garssen (eds.) *Controversy and Confrontation: Relating Controversy Analysis with Argumentation Theory.* Amsterdam and Philadelphia: John Benjamins. 109-124.

Gotti, M. (2006), 'Disseminating early modern science: specialized news discourse in the Philosophical Transactions', in: N. Brownlees (ed.) *News Discourse in Early Modern Britain: Selected Papers of CHINED 2004*. Bern: Peter Lang. 41-70.

Grmek M.D. (1998), *Western Medical Thought from Antiquity to the Middle Ages.* Cambridge, Massachusetts and London: Harvard University Press.

Jones, C. (2004), 'Discourse communities and medical texts', in: I. Taavitsainen and P. Pahta (eds.) *Medical and Scientific Writing in Late Medieval English*. Cambridge: Cambridge University Press. 23-36.

Mäkinen, M. (2006), *Between Herbals et alia: Intertextuality in Medieval English Herbals*. PhD thesis, University of Helsinki. <http://ethesis.helsinki.fi/julkaisut/hum/engla/vk/makinen/>.

Margolies, D. (1985), *Novel and Society in Elizabethan England.* London and Sydney: Croom Helm.

McEnery, T. (2006), 'The moral panic about bad language in England 1691-1745', *Journal of Historical Pragmatics*, (7)1: 89-113.

Minnis, A.J. (1979), 'Late medieval discussions of *compilatio* and the rôle of the *compilator*', *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 101: 385-421.

Mössner, L. (forthcoming), 'The influence of the Royal Society on 17th-century scientific writing', *ICAME Journal*.

Ratia, M. (forthcoming), *Argumentative Strategies in the Early Modern Tobacco Controversy*. PhD thesis, University of Helsinki.

Shapin, S. (1996), *The Scientific Revolution.* Chicago and London: The University of Chicago Press.

Siraisi, N. (1990), *Medieval and Early Renaissance Medicine: An Introduction to Knowledge and Practice*. Chicago and London: University of Chicago Press.

Swales, J.M. (1990), *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Taavitsainen, I. (1993), 'Genre/subgenre styles in Late Middle English?', in: M. Rissanen, M. Kytö and M. Palander-Collin (eds.) *Early English in the Computer Age.* Berlin and New York: Mouton de Gruyter. 171-200.

Taavitsainen, I. (2001a), 'Changing conventions of writing: the dynamics of genres, text types, and text traditions', *European Journal of English Studies*, special issue, edited by L. Mössner: 139-150.

Taavitsainen, I. (2001b), 'Evidentiality and scientific thought-styles: English medical writing in Late Middle English and Early Modern English', in: M. Gotti and M. Dossena (eds.) *Modality in Specialized Texts: Selected Papers of the 1st CERLIS Conference*. Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien: Peter Lang. 21-52.

Taavitsainen, I. (2002a), 'Variation and formularity: prefatory materials in early scientific writing', in: K. Lentz and R. Möhlig (eds.) *Of Dyuersite & Chaunge of Langage: Essays Presented to Manfred Görlach on the Occa-*

sion of his 65th Birthday. Heidelberg: C. Winter Universitätsverlag. 290-304.

Taavitsainen, I. (2002b), 'Historical discourse analysis: scientific language and changing thought-styles', in: T. Fanego, B. Méndez-Naya and E. Seoane (eds.) *Sounds, Words, Texts, Change. Selected Papers from the Eleventh International Conference on English Historical Linguistics.* Amsterdam and Philadelphia: John Benjamins. 201-226.

Taavitsainen, I. (2004), 'Transferring classical discourse conventions into the vernacular', in: I. Taavitsainen and P. Pahta (eds.) *Medical and Scientific Writing in Late Medieval English*. Cambridge: Cambridge University Press. 37-72.

Taavitsainen, I. (2005), 'Genres and the appropriation of science: *loci communes* in English literature in late medieval and early modern periods', in: J. Skaffari, M. Peikola, R. Carroll, R. Hiltunen and B. Wårvik (eds.) *Opening Windows on Texts and Discourses of the Past*. Amsterdam and Philadelphia: John Benjamins. 179-196.

Taavitsainen, I. (forthcoming a), 'Dissemination and appropriation of medical ideas', in: I. Taavitsainen and P. Pahta (eds.) *Medical Writing in Early Modern English*. Cambridge: Cambridge University Press.

Taavitsainen, I. (forthcoming b), '"Joyful news out of the newfound world": medical and scientific news reports in Early Modern England', in: A.H. Jucker (ed.) *Early Modern English News Discourse: Newspapers, Pamphlets and Scientific News Discourse.* Amsterdam and Philadelphia: John Benjamins.

Taavitsainen, I. (forthcoming c), "Authority and instruction in two sixteenth-century medical dialogues", in: M. Peikola, J. Skaffari and S-K. Tanskanen (eds.) *Instructional Writing in English: Studies in Honour of Risto Hiltunen.* Amsterdam and Philadelphia: John Benjamins.

Taavitsainen I. and S. Fitzmaurice (2007), 'Historical pragmatics – what it is and how to do it?', in: S. Fitzmaurice and I. Taavitsainen (eds.) *Methods in Historical Pragmatics.* Berlin and New York: Mouton de Gruyter. 11-36.

Taavitsainen, I. and P. Pahta (1998), 'Vernacularisation of medical writing in English: a corpus-based study of scholasticism', *Early Science and Medicine*, special issue, edited by W. Crossgrove, M. Schleissner and L.E. Voigts: 157-185.

Valle, E. (1999), *A Collective Intelligence. The Life Sciences in the Royal Society as a Scientific Discourse Community, 1665-1965*. Anglicana Turkuensia 17. University of Turku.

Valle, E. (2006), 'Reporting the doings of the curious: authors and editors in the Philosophical Transactions of the Royal Society of London', in: N. Brownlees (ed.) *News Discourse in Early Modern Britain. Selected Papers of CHINED 2004.* Bern: Peter Lang. 71-90.

Voigts, L.E. (1984), 'Medical prose', in: A.S.G. Edwards (ed.) *Middle English Prose: A Critical Guide to Major Authors and Genres*. New Brunswick, New Jersey: Rutgers University Press. 315-335.

Wear, A. (2000), *Knowledge and Practice in English Medicine, 1550-1680*. Cambridge: Cambridge University Press.

A diachronic perspective on changing routines in texts[1]

Tanja Rütten

University of Cologne

Abstract

This paper explores the role of exhortation in religious discourse and is intended as a case study of higher-level pragmatic analyses of texts by way of corpora and corpus methodology. Whereas diachronic pragmatic descriptions usually focus on the development of single (and isolated) items to characterize the history of a text or genre, this study accesses texts on the level of entire subsections. A diachronic outline of exhortation as one such section in religious instructive discourse shows that it undergoes a process of dissolution and integration: it falls out of use as a complex component of instructive discourse in favour of smaller, more integrated elements of exhortation. A diachronic, contrastive view of exhortation in sermons and treatises shows that this development is sensitive to time as well as genre.

1. Introduction

Diachronic pragmatic studies usually access texts or genres by analysing isolated features or a combination of single features, which may typically be located on the level of sentence or below. Approaches vary from multidimensional feature analysis (Biber and Finegan 1992) to the analysis of speech acts, discourse markers or linguistic manifestations of politeness (e.g. in address term systems) by which the pragmatic history of texts and genres is characterized (e.g. Taavitsainen and Pahta 1995; Arnovick 1999; Taavitsainen and Jucker 2003; Walker 2003). Corpora and corpus methodology have served in various ways in many of these studies. However, texts are complex constructs and it seems difficult to assess them accurately on the basis of single elements. Kohnen argues that a text is composed of "higher-level functions and sections" which are determined by the functional profile of the genre to which the text belongs (see Kohnen, this volume and 2007a). Since such sections may constitute valuable links between genres and may reveal an underlying coherent network, Kohnen suggests including them in a diachronic description of texts and genres. This study will show how such an approach may contribute to a more comprehensive picture of the pragmatic history of texts: the development of religious instructive discourse is assessed here on the basis of its most crucial component, that is, exhortation. It will be shown that exhortation disintegrates as a complex functional section and that it is realized instead in the shape of single, flexible

elements in instructive discourse. This study is intended as a case study and is based on a corpus of 135,000 words, covering the two key genres of religious instruction, sermons and treatises, from the late fourteenth to the seventeenth century (for details see appendix).

Exhortation is one of five recurring subfunctions within religious instructive discourse, next to exposition, exegesis, narration and argumentation (cf. Kohnen 2007a; see also Werlich 1976 for a description of basic textual functions). Each of these functions occurs in a text in changing proportions and in various forms, and is shaped by the needs of the discourse community. Exegesis, for example, is a typical feature of medieval instructive discourse, mainly found in homilies and sermons.[2] Its importance diminishes during the Reformation, but it is resumed as a pattern of instruction at the turn of the sixteenth century, motivated by the establishment of Protestantism and its emphasis on the written word. Although, in a way, this is a return to more traditional means of instruction, it is now usually found in Bible commentaries, i.e. in exegetical treatises, rather than in sermons. Thus, exegesis undergoes a twofold shift, one in terms of varying prominence in the discourse community throughout time and one in terms of genre, from sermon to treatise (and thus, from a more oral to a rather literate form of communication).[3] It supports the fact that the above-mentioned textual functions are variable means of instruction, that their occurrence is motivated by the needs of the language user and that, subsequently, their status in the discourse world may be subject to change.

To come to an understanding of how texts work in a discourse community and how they evolve, it is thus quite instructive to access them through such functional sections. As the example of exegesis shows, it seems difficult to point out preferred ways of combining these functions in a text off-hand or to make claims on their prominence at a given point in time, let alone their development throughout time and in various genres. But even though they combine in an "unforeseeable way" (Kohnen 2007a), they create unique patterns which, once identified, may be commonly associated with a particular text or genre at a specific point in time. Eventually, this will allow us to make more confident statements on routines which are followed, abandoned or even newly created in texts and genres.

In analysing textual subsections and functions one needs to consider the fact that they are complex entities. They usually cover larger portions of a text, are demarcated from other sections in various ways and consist of a sequence of (more or less typical) individual illocutions which, taken together, serve a common purpose. They cannot be captured in their entirety by an analysis of single speech acts alone but present higher-level functional elements of which a speech act is the most basic element. A short characterization of exhortation will serve to demonstrate this functional and structural complexity.[4]

2. The role of exhortation in religious instruction: a case study

2.1 Exhortation as a pattern of religious instruction

Exhortation may be considered the core function of religious instruction and is the most direct and effective means on the side of the church to regulate believers' physical and mental conduct, telling them what to do and think, respectively. Its ultimate aim is to make the individual believer conform to the doctrine (with varying degrees of imposition of this regulation upon him/her, ranging from direct commands to rather polite requests or suggestions). In the present data, this aim is represented in two distinct ways: it occurs as a higher-level component as well as in the shape of single and isolated speech acts in a text; whereas the first is independent and structurally and illocutionary complex, the latter is quite simple in terms of structure and function, and is clearly subordinate to other sections/functions. Consider the following two examples (italics mine in all examples):

(1) þys was þe cause of Crystys fyrst comyng ynto þys world. *Wherfor* he þat wyll scape þe dome þat he wyll come to at þe second comyng, *he most lay downe* all maner of pride and heynes of hert, and *know hymselfe þat* he ys not but a wryche and slyme of erth, and soo *hold mekenes yn his hert*. *He most trauayl* his body yn good werkes, and *gete* his lyfe wyth swynke, and *put away* all ydylnes and slewth. For he þat wyll not trauayle here wyth men, *as Seynt Barnard sayth*, he schall trauayle ay wyth þe fendes of hell. And for dred of deth *he mot make hym redy to his God*, […]. *þen schall he haue* yn þe day of dome gret remedy and worschip. […] *And þose þat haue not schryuen hom*, hit schall be schowet to all þe world yn gret confusyon and schenschyp. *Þys ys sayde for þe fyrst commyng of Cryst ynto þys world*. (John Mirk, Festial, *Sermon for Advent Sunday*, a1415)

Example (1) is a typical section dedicated to exhortation. It instructs the community of believers how to lead a good, Christian life, and how to avoid punishment on Doomsday. It is a section distinct from its (expository) context both in terms of structure and function. As concerns structure, its boundaries are clearly outlined by discourse markers:[5] the exposition concludes with a closing statement ("þys was þe cause of Crystys fyrst comyng ynto þys world") relating to and summarizing the previous section. The exhortation then opens with the introductory marker "wherfor" and is indicated by a change in tense as well. A concluding remark summarizes and links the exhortation to the expository context ("Þys ys sayde for þe fyrst commyng of Cryst ynto þys world"); structurally, this is a parallel to the concluding statement in the expository section. The text then continues with an exposition of the second of the two comings of Christ, a concept originally introduced at the very beginning of the discourse.

As concerns function, a shared purpose and interrelation of the individual illocutions in this section is quite evident: a sequence of explicit regulations (i.e. directive acts) is followed by a presentation of the resulting profits and a statement on the consequences if the believer fails to comply with these regulations. Church authorities, such as the Church Fathers or a saint, are quoted to give more weight to the regulations.

Usually, this kind of exhortation is global in scope, regulating the moral conduct of the audience much beyond the actual discourse, offering advice or giving orders for proper behaviour to be followed quite generally in life (here: what to do to escape punishment on Doomsday).

Example (2), on the other hand, is only a single regulation, embedded in an otherwise exegetical context:

(2) Good men and wymmen, oure Lorde Ihesu tauȝth is disciples, as þe gospell wittenesseþ, (Mathei 6to, Luce 11mo) þis preyoure of þe Pater Noster, *þe wiche þat euery man shuld preye to God* when þat þei preyed, as Poule dude by þe wordes of my teme, þus seynge: "I do þonkes to my Lorde God," as I seid at þe begynnynge. In þis worthy prayere of þe Pater Noster ben vij asshyngus, þe wiche iij firste perteynen to þe þre persons in Trynite and oo God. (anon., *Middle English Sermons*, 1400-1415)

Here, the regulation functions more as an addition to the interpretation of the Lord's Prayer than as a constitutive text component. Its most striking characteristic is its syntactic integration: it occurs in a non-defining relative clause modifying the noun phrase "þe Pater Noster". This syntactic subordination makes the regulation 'to pray to God by the Pater Noster' an integral part of the exegesis and implies its ancillary position in the discourse. It shows little of the structural and illocutionary complexity observed in (1), only a short reference to St. Paul as an authority is given.

Early Modern examples are often even more confined syntactically, see for instance (3), where the regulation is restricted to the apodosis in a conditional sentence. Here, the exhortation works rather locally in the discourse and is firmly embedded in an argumentative context:

(3) If yee desire yet farther to knowe how necessarie and needfull it is, that we edifie and build vp our selues in faith, *marke the words of the blessed Apostles*, 'without faith it is impossible to please God'. (Richard Hooker, *Two Sermons upon Part of S. Jude's Epistle*, 1614)

Even though such representations are rather simple in terms of structure and function and only have a local scope, they serve the same exhortative purpose, nevertheless; such single acts are referred to as 'exhortative acts', whereas the complex pattern is referred to as 'exhortation proper'.

Concerning linguistic realizations, both exhortation proper and exhortative acts make ample use of imperatives and modal verb constructions. Single

exhortative acts, however, often seem less imposing than the straightforward pattern proper in example (1), which is conspicuous by the sheer mass of regulations alone and strongly stands out as a distinct section of text.

To make quantifiable statements on the distribution and development of both exhortation proper and exhortative acts, an analysis was conducted with parts of the *Corpus of English Religious Prose* (see appendix). Text functions were coded manually and measured by relative frequency of occurrence as well as relative frequency of words dedicated to each. Relevant passages were identified by a combined focus on function, illocutionary coherence, discourse markers and cohesive means, as illustrated in examples (1) to (3). To arrive at a more comprehensive understanding of representations of exhortation in a given text and genre, coding included all five textual functions (i.e. exhortation, exposition, exegesis, narration and argumentation). The analysis shows that both kinds of exhortation occur in the data in synchronic as well as in diachronic variation, and that their occurrence may be linked to both temporal and genre-specific constraints.

2.2 The diachronic development of exhortation

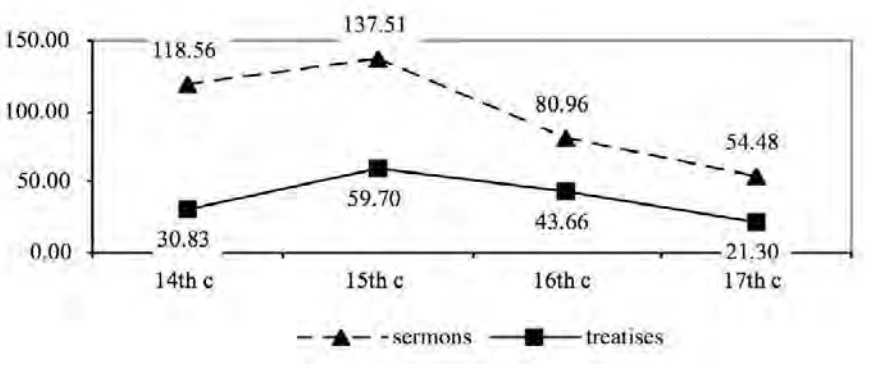

Figure 1: Proportions of exhortation in sermons and treatises from the fourteenth to the seventeenth century (number of words dedicated to exhortation per 1,000 words)

Figure 1 shows the average amount of words dedicated to exhortation for both genres under investigation from the fourteenth to the seventeenth century (both representations, i.e. complex text components and single acts, are combined).[6] Generally, exhortation as a means of instruction is on the decline in both genres with an average volume of 118.5/1,000 words in the fourteenth century and 54.4/1,000 words in the seventeenth century. There is a significant decrease at the turn of the sixteenth century and considerably more exhortation in Middle English than in Early Modern English (former: 95.8/1,000 words; latter: 50.0/1,000 words), rendering exhortation a more medieval kind of instruction.

That this change is not merely one of size, i.e. a development towards a more economic realization of exhortation, but one in actual decrease of frequency is confirmed by table 1, which lists the frequency of occurrence of exhortation in both genres:

Table 1: The occurrence of exhortation in sermons and treatises from the fourteenth to the seventeenth century (frequency per 10,000 words; numbers in brackets indicate absolute frequencies)

		14th c	15th c	16th c	17th c
Sermons	Exhortation proper	3.8 [4]	4.5 [9]	1.0 [2]	0.5 [1]
	Exhortative acts	32.9 [34]	13.6 [27]	35.4 [71]	21.3 [43]
Treatises	Exhortation proper	0.0	0.0	0.0	0.0
	Exhortative acts	17.9 [9]	24.4 [49]	22.1 [43]	11.3 [23]

Correlation values (*Pearson product-moment correlation coefficient*) between size and frequency are r=0.30 for sermons and r=0.98 for treatises, indicating that exhortation declines both in terms of size and frequency in both genres. However, whereas treatises show a notable uniformity in this development, the correlation is less ideal in sermons. Here, we find an increase in size coupled with a decrease in frequency in the fifteenth century (compare figure 1 and table 1); this development can be interpreted as a sign of genre-internal variation and is addressed in section 2.3

The importance of exhortation as a means of instruction in Late Middle English observed here may be linked to the Fourth Lateran Council of 1215, which, according to Barratt, shaped religious instructive discourse for centuries (Barratt 1984: 413). Constitutions on yearly confession, one of the decisive outcomes of the Council, necessitated knowledge about what constituted sin in the first place and therefore knowledge about the Ten Commandments, how to remedy sinful behaviour and prepare for Doomsday, or how to approach God in prayer and ask for forgiveness of one's sins. Both sermons and treatises reflect this 'teaching plan' and show how rigid and fixed religious instruction was for much of the later Middle Ages. Exhortation was a crucial means to meet the requirements of Lateran IV. It generated a substantial sum of regulations of a quite general nature and much of the instructive discourse at the time was characterized by passages such as these:

(4) This ten Comandementh that I haue nowe rekend Er umbilouked in twa of the godspell, (Luce x°. cap°.) The tane is we love god ouer al thinges, The tothir that we love our euen-cristen als we do oure selven. For *god augh us to love halye with hert*, With al our might, with al our thought, with word and with deid: *Our euen-cristen alswa augh us to loue Un-to that ilk gode that we loue us selven*, That is, that thai welefare in bodi and in saule, And cum to that ilk blisse that we think to. Who-so dos this twa fulfilles the othir. (anon., *The Lay Folk's Catechism*, 1373)

(5) þis prayore *euery Cristen man is bondon to conne*, and *to preye to God by þis preyour*, and *do as Seynt Poule dothe* by þe wordes of my teme: "I do þonkes to my Lord God." *þis prayoure is euery man and childe hold to kunne* ȝif he passe vij ȝere olde; and þere frendes be in grett perill ȝeff hei teche hem nott is Pater Noster, Aue Maria, and is Beleue. (anon., *Middle English Sermons*, 1400-1415)

The general decline of exhortation observed in the data, in particular in the sixteenth century, testifies to massive changes in instructive discourse. The Reformation challenged many of the traditional views and customs and it became necessary for orthodox and reformist clerics alike to argue with and convince their audience, or at any rate, to show greater consideration and awareness for their needs if they wanted to be effective. An audience unsettled by new interpretations of the doctrine and new customs, and one which began to question orthodox theology, could no longer simply be told what to think or do, but needed to be reassured of their belief and convinced of the excellence of their faith and clerics had to react upon this development.

Subsequently, exhortation was obviously no longer considered the most appropriate means to meet these ends in religious instruction. Where it was still used, however, its linguistic inventory shows that regulations were now frequently put in a form which was much less imposing on the recipient, which also shows that authors were aware of the changing needs of their audience. Whereas Middle English authors preferred direct and explicit forms (see example (5)), the illocutionary force of the regulations in Early Modern English was mitigated in two ways (see examples (6) and (7)): Early Modern English authors prefer verbs denoting cognition such as 'observe' or 'consider' in the regulation which lay it open to the consideration of the audience, clearly less obliging than the forms chosen in (4) or (5). Authors also mitigate a direct regulation by means of a less direct construction. This is shown in (7), a popular sixteenth century exegetical commentary, where the author includes himself in his regulation, thereby weakening the illocutionary force of the act:

(6) *Observe* then, That it is no new thing, if you find in God's Church, barren Fig-trees, fruitless Professors. (John Bunyan, *The Barren Fig Tree*, 1673)

(7) And thairfoir, albeit we be laid open sumtymes, as it wer, evin to the mouth of Sathan, *lat us not think thairfoir that God hath abjectit us*, and that he takith no cair over us. (John Knox, *An Exposition upon Matthew IV*, 1556)[7]

The observation that exhortation as a pattern of instruction is on the decline and that its appearance in Early Modern English shows signs of 'audience design' (cf. Bell 1984) is in line with findings from studies which, for instance, have looked at the development of directives (e.g. Kohnen 2007b) or studies on the history of politeness (e.g. Jucker 2008). In general, both claim that there is a tendency towards less direct forms of directives, less directive, face-threatening acts in

general and a tendency to use more negative politeness in Early Modern English – claims which clearly argue for less exhortation (an inherently face-threatening activity) and more consideration of the recipient in instructive discourse.[8] Contrasting both exhortation proper and exhortative acts individually reveals a more detailed picture of this process (see figure 2).

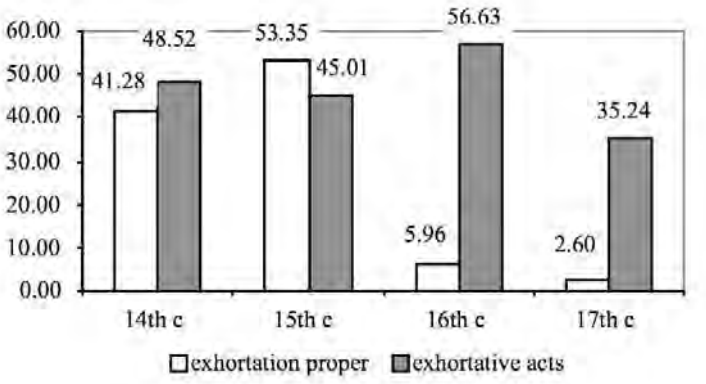

Figure 2: Proportions of exhortation proper and exhortative acts in both genres from the fourteenth to the seventeenth century (number of words dedicated to each per 1,000 words)

In Middle English, figures for exhortation proper and exhortative acts are rather even: both kinds of exhortation occur in similar proportions in the data. Early Modern English texts, on the other hand, almost exclusively make use of exhortative acts in instructive discourse, abandoning the pattern proper almost entirely.

The decline in the sixteenth century observed in figure 1 now appears in a different light: it first and foremost concerns the pattern proper and is quite drastic (fifteenth century: 53.3/1,000 words; sixteenth century: 5.9/1,000 words). Figures for exhortative acts, in contrast, are surprisingly stable and continually high. In the sixteenth century they actually rise, a development which is likely to be the result of the decline of exhortation proper in the same period.[9] In the seventeenth century, exhortative acts are almost used exclusively in the data and still occur with considerable frequency (seventeenth century sermons: 21.3/10,000 words; seventeenth century treatises: 11.3/10,000 words; cf. table 1).

What this seems to suggest is that exhortation as a complex text component dissolves and that it is increasingly integrated into the instructive discourse by means of single regulations. Whereas in Middle English, both kinds occur in synchronic variation, in Early Modern English, exhortation comes to be realized in isolated directive speech acts embedded in other contexts. Exhortation as a complex pattern falls out of use completely and what remains of exhortation is much more flexible in the discourse since it occurs in a text wherever necessary to direct the recipient's behaviour. In other words, exhortative acts enter other

text sections (i.e exposition, exegesis, narration or argumentation) and become integral parts of them. Figure 3 shows in which of the other sections exhortative acts are most likely to be found in the data and identifies preferred patterns. Note that in Middle English argumentation did not exist as a distinct functional section.[10]

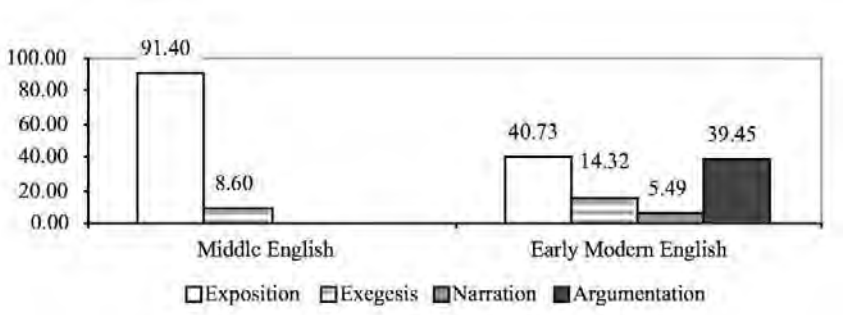

Figure 3: Distribution of exhortative acts across other sections in both sermons and treatises (figures in percent)

In Middle English, the overwhelming majority of all exhortative acts is found in exposition, but some also occur in exegesis. There is a marked absence of exhortative acts in narration, a section which may consequently be considered an end in itself. This absence seems plausible from what is known about the tradition of narratives in religious writing. Medieval narratives were typically intended to simply impart Biblical contents or, as in the case of exempla, were used more for entertainment than for learning (cf. Day 1974; Kahrl 1966). At a time of overwhelming illiteracy and when only single attempts at Englishing the Bible had been made (cf. Green 2000), it was a major concern of the clergy to simply pass on knowledge of Biblical contents. The lack of any subsequent exhortation following the narrative in example (8) seems to substantiate the view that no didactic purpose was intended except for the presentation of the actual content.

(8) Ihesus was ledde into desert of a spirit, þat he shuld be temptid of þe deuel. And whan he had fastid fourti daies and fourti nyȝtis, aftirward he hungrid. And þe tempter, þe feend, come to hym, and seide, "If þou be Goddis Sone, commaunde þat þese stoones be maad looues." And he answerid hym and seide, "It is wryten: 'Not oonli in breed lyueþ man, but also in euery word þat goþ forþ of Goddis mouþ.'" þþan þe deuel took hym ȝinto þe holi citee, Ierusalem, and set hym aboue þe pynnacle of þe temple. And he seide to hym, "If þou be þe Sone of God, late þisilf dounward; for whi, it is writen þat: 'God þe Fadir haþ commaundid to His angels of þee, and þei shuln take þee in her hondis, lest þou stomble þi foot at a stoon.'" And Ihesus seide to hym eftsones, "It is writen: "þþou shalt not tempte þi Lord God."" Eftsones þe deuel took hym into a ful hyȝ hul, and shewid hym alle þe rewmes of þe world, and þe glorie of hem. And he

> seide to hym, "Alle þese þyngs I shal ȝeue to þee, if þou, fallyng doun, worship me." þþanne Ihesus seide to hym, "Go, Satanas: For it is writen: 'þþou shalt worshipe þi Lord God, and Hym aloone þou shalt serue.'" þþanne þe feend leeft hym, and lo, aungels comen to, and mynystreden hym. (anon., Lenten Sermons, *Sermon for the Second Sunday in Lent*, a1400)

Early Modern narratives often tend to be personal reminiscences or 'private' stories by the cleric rather than Biblical stories, even though those still exist. Thus, an explanation of the lesson to be learned from the story *qua* exhortation seems necessary:

(9) But *horsses are not to be preferred aboue pore men*. I was ones offendid with the kynges horses, and therfore toke occasion to speake in the presens of the kynges maiestye that dead is, whan Abbeis stode. Abbeis were ordeyned for the comforte of the pore, Wherfore I sayde it was not decent that the kings horsses shuld be kept in them (as manye were at that tyme) the lyuynge of poore men therby minished and taken a way: But after ward a certayne noble man sayd to me, what hast thou to do with ye kinges horsses? I answered, and sayd, I speake my conscience as goddes word directeth me. He said horsses be ye mayntenaunces and parte of a kynges honoure, and also of hys realme, wherefore in speakyng againste them ye are against the kynges honoure. I answered. God teacheth what honoure is decente for the kynge and for al other men accordynge vnto their vocations. God apoynteth euery king a sufficient lyuinge for hys state and degre boeth by landes and other customes. And it is lawfull for euery kyng to enioye the same goodes and possessyons. But to extorte and take awaye the ryghte of the poore, is agaynste the honoure of the kinge. And you do moue the kinge to do after that manner, then you speake agaynste the honoure of the kynge. (Hugh Latimer, *Seven Sermons before Edward VI*, 1549)

In preaching against the exceeding riches of a king, Latimer relates a conversation he had had with a 'certain noble man'. The purpose of this story is clearly formulated at the outset. Latimer instructs his audience (the king and his court, and thus a rather delicate communicative situation) that horses are not to be preferred over men. The story is similar in length to the one in (8), but the content is quite different and serves a completely different purpose. Example (8) is intended to impart knowledge of the three temptations of Jesus by the Devil, a popular medieval topic in Lent. The story in example (9), on the other hand, is a clever rhetorical device to voice arguments against the luxury of the king and his court without offending them by addressing either of them directly. The regulation used in combination with the story clarifies its purpose.

Contrasting Late Middle and Early Modern English sermons and treatises, a more even distribution of exhortative acts across other sections can be seen in

figure 3 (above). In particular, there is a strong similarity between their patterning with exposition and argumentation, which corroborates the claim that argumentation is a particular kind of exposition or at least strongly related to it (see Werlich's description of the 'expository' and 'argumentative text idiom'; 1976: 256-265). Exhortation is used in both cases to summarize a point or an argument, to introduce a new idea or drive home a point. Note how similar in structure both (10) and (11) are:

(10) WHAT PRAYER IS.- *Who will pray, must knowe and understand that Prayer is ane earnest and familiar talking with God*, to whome we declair oure misereis, whois support and help we implore and desyre in our adversiteis, and whome we laude and prais for oure benefittis receaved. So that Prayer conteaneth the expositioun of our dolouris, the desyre of Godis defence, and the praising of his magnificent name, as the Psalmis of David cleirlie do teache. (John Knox, *A Declaration*, 1554)

(10) is an example of a patterning of exhortation with exposition. The audience is first exhorted to understand what the nature of prayer is before the purpose of a prayer is expounded upon in detail. In (11), the audience is asked to reflect upon the argumentation made thus far in the text to confute Martin Luther's reformist views. The exhortation here serves to sum up the line of argumentation and to entice the audience to reconsider the whole issue:

(11) *Consyder now how eche of these testymonyes conferme & strengthe one another*. Fyrste the fygure & shadow of the olde lawe. Secondly the testymony of the gospels answerynge vnto the same. Thirdly the declaracyon of saynt Austyn vpon the same. And here I brynge but one doctour. whose testymony in the balaunce on any trewe christen mans herte. me thynketh sholde weye downe Martyn Luther. (John Fisher, *Sermon against Luther*, 1521)[11]

Exhortation in Early Modern English is thus no longer a fixed and complex component in texts but is a flexible means of instruction, typically working on a local level as the two examples above show.

2.3 Genre constraints on exhortation

Contrasting the use of exhortation in both genres, it becomes clear that the dissolution of the pattern is strongly genre-sensitive (see figure 4).

What is most striking is the complete absence of exhortation proper from treatises. Here, exhortation is only represented in the form of single acts. However, it should be mentioned that treatises are a versatile and markedly multifunctional genre. Much writing, especially in Early Modern English, has been termed 'treatise' indiscriminately (cf. Green 2000), and it is necessary to draw finer distinctions. One such distinction could be to differentiate doctrinal treatises (dealing with basic points of the doctrine, e.g. the Decalogue, the Pater

Noster, or the Seven Deadly Sins) from controversial treatises (cf. Green 2000: 216ff) and contemplative treatises (dealing with higher forms of spirituality and religious experience, such as those in the mystic tradition (but see also Sargent 1984)). This study only includes doctrinal treatises, but a tentative analysis of some of the major contemplative treatises has shown that exhortation proper occurs in contemplative writing in proportions equal to those of sermons (the fourteenth-century texts *The Abbey of the Holy Ghost*, *The Epistle of the Discression of the Spirit* and the famous *Cloud of Unknowing* yielded a relative frequency of words of 33.6, 214.7 and 95.4 of exhortation proper per 1,000 words, respectively).

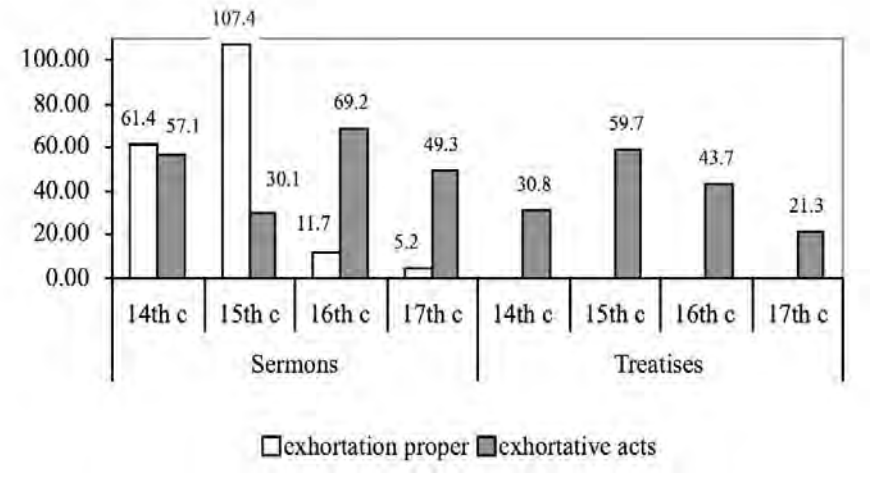

Figure 4: Proportions of exhortation proper and exhortative acts in sermons and treatises (number of words dedicated to each per 1,000 words)

Contrasting doctrinal treatises and sermons, however, suggests that exhortation proper seems to have been much more typical of sermons. Assuming that what has come down to us as a 'sermon' (necessarily a written text) is based on an oral communicative situation, it seems feasible to assume that here a priest would be most inclined to influence the moral behaviour of his flock and make frequent use of exhortation.

Another development which seems genre-sensitive is that of exhortation proper in the fifteenth century: sermons show a strong increase of the pattern proper in this period. This may be interpreted as a sign of genre-internal variation which is caused by model sermons and cycle collections. Both were extremely popular in the fifteenth century and follow rather strict and rigid routines.[12] Composed (or compiled) to be re-used and recycled year after year, they could be

utilized for different audiences and at different places. Their functional profile, subsequently, needs to permit this variability, encouraging regulations with a general validity, prefabricated with references to authorities, assurance of either salvation or damnation and strongly discouraging individuality as concerns the regulations for a largely unknown and indeterminate audience. All this may render exhortation proper more appropriate for this subgenre. Example (12) is taken from a model sermon. It offers a general definition of who is exempt from fasting (children and old people, pregnant women, pilgrims etc.) and is quite vague about the audience it addresses ("yche man", "ye"), the place ("com to holy chyrche") and the member of the clergy ("þe prest"). This makes the sermon, and the regulations contained in it, quite suitable to be preached anew each year:

(12) And for ych man dothe forfet more oþer las, þerfor, forto make satysfaccyon for þat gylt, *yche man ys holden* by þe lawe of holy chyrch to fast þes fourty dayes, outtaken hom þat þe lawe dyspensyth wyth for nede. That ben chyldyr wythyn xxi here, woymen wyth chyld, old men passed age and myghtles to fast, pylgrymys, and seke, and pore, and þes þat labryn sore fore hor lyuelod: þus þe lawe dyspensyth wyth apon hore concyens. þen for bycause þat Sonday ys no day of fastyng, þerfor *ȝe schull begyn your fast at Aske-Wanysday*, *and þat day com to holy chyrche*, *and take askes at þe prestys hond*, *and ber forth in your hert þat he sayth to you*, when he layth askes on your hedys. þen he saythe þus: 'Man, thynke þat þou art but eskys, and to eskys þou schalt aȝeyne turne'. (John Mirk, Festial, *de Dominica in Quadragesima*, a1415)

The example of a later, single sermon in (13), in contrast, is much more precise in the address to the audience; the content has clearly been delivered for one single occasion and can hardly be re-used (except for private reading at home, that is). The regulations concerning the citizens of London are quite specific and several times individual groups are singled out and addressed directly ('ye that be prelates', 'lordes that live like loiterers', 'you that have charge of youth'):

(13) *Oh London London, repente repente*, […] Amende therfore and *ye that be prelates loke well to your office*, for right prelatynge is busye labourynge and not lordyng. Therfore preache and teach and let your ploughe be doynge, *ye lordes I saye that liue lyke loyterers, loke well to your office,* the ploughe is your office and charge. If you lyue idle and loyter, you do not your duetie, you folowe not youre vocation, let your plough therfore be going and not cease, that the ground maye brynge foorth fruite […] Therefore for the loue of God, *appoynte teachers and scholemaisters, you that haue charge of youth*, and giue the teachers stipendes worthy their paynes, that they maye brynge them vp in grammer, in Logike, in rethorike, in Philosophe, in in the ciuile lawe, and in that whiche I can not

> leaue vnspoken of, the word of God. (Hugh Latimer, *Sermon on the Ploughers*, 1549)

We see here that exhortation is strongly sensitive to communicative context as reflected in genres and subgenres: exhortation proper is most typically found in sermon collections and relatively oral contexts. Doctrinal treatises as systematic, methodical written texts do well without complex exhortative patterns and prefer single, hence flexible, regulations.

3. Changing routines in religious instruction

Having briefly sketched a history of exhortation in two core genres, it has become clear that it is a pattern of instruction subject to considerable change. The evolution from a pattern proper to an integrated element in other sections of instructive texts affects single genres and subgenres to different degrees and may even be a feature discriminating them. Three things have become apparent in this study:

1. Exhortation is on a steady decline, but whereas the pattern proper falls out of use completely as a means of instruction, exhortative acts proliferate rather unaffectedly, despite the general tendency towards less exhortation. By the seventeenth century, they have become the typical representation of exhortation in instructive discourse.
2. This observed change in routines is a rather sudden, abrupt and far-reaching development at the beginning of the Early Modern English period. However, there is also stability observed in this process: exhortation is not simply abandoned as a pattern of instruction but evolves into a more discreet and subtle means, surfacing in any of the other sections in a text, thereby changing the appearance of such sections to a considerable extent (see, for instance, the discussion of narration). I have argued that strategic considerations of face and politeness may serve as possible explanations for this change but stylistic reasons may be responsible for the rejection of exhortation proper as well.
3. Exhortation is strongly genre-sensitive: different genres and subgenres show a greater or lesser liking for the pattern, as we saw in the case of single and cycle sermons or doctrinal and contemplative treatises, for example. Here, the underlying communicative situation, in particular the reception of the text by the reading or listening audience, may be a major factor influencing the choice.

On a more general level, this study has shown that genres and subgenres alike follow quite distinct routines. A diachronic analysis extended to each of the five subsections and additional genres, brings to light such routines and seems quite useful as a tool to establish more comprehensive diachronic descriptions of texts. This study has highlighted the importance of analysing the history of genres on a level higher than that of single speech acts. A textual history based on an outline

of single illocutions alone hides important developments and equals larger textual structures and even complete texts to the development of the very speech acts of which they are created. The discussion of the development of both exhortation proper and exhortative acts is a case in point.

I would like to conclude with the observation that the role of exhortation in instructive discourse is preliminary without an assessment of catechisms in this development. This is particularly so because catechisms begin to occur on the scene at exactly the time when the changes discussed in this study are observed (cf. Green 1996). Suffice it here to illustrate the exhortative tone of catechisms with two examples (one where the regulations become apparent in the questions and one where they occur in the answers):

(14) *Question.* You sayde that your Godfathers and Godmothers dyd promyse for you that ye should kepe Goddes commaundementes. *Tell me how many there bee.*
Aunswere. Tenne.
Question. Whiche be they?
Aunswere. […]
Question. What dooest thou chiefely learne by these commaundementes?
Aunswere. I learne two thinges: My duetie towardes god, and my duetie towardes my neighbour.
Question. *What is thy duetie towardes god*? (anon., *The Catechism contained in the Book of Common Prayer*, 1549)

(15) *Father*. How ought ministers to deale with many ignorant & simple men, & seely soules which are able to yeeld small reason of their faith: are they all to be admitted to the communion, or all to be reiected?
Child. As *ministers ought not rashly to accept of all that offer themselues without examination and conference: so ought they not lightly to shut out any from that which should seale vp their remission of sinnes*. For albeit they be somewhat ignorant and simple, yet finding in them any seeds of religion, […] they are to admit them with encouraging and exhorting them to go foreward. But if with ignorance and blindnesse, be ioyned either froward contempt of meanes, or meere carelesnesse, or open wicked behauiour, or profane dissolutenes without remorse when they be dealt withall; *Then they are with all mildnesse to be shut out.* All glory be giuen to God. (Arthur Dent, *A Pastime for Parents*, 1606)

Catechisms are an important means in Early Modern religious instruction for the "understanding of the simple" (cf. Dent 1601, title page), and an assessment of religious instruction is incomplete without an evaluation of their role in this process.

Notes

1 I would like to thank Thomas Kohnen for his valuable comments on an earlier draft of this manuscript.

2 On the distinction of both genres and their eventual fusion during Late Middle English see Heffernan (1984) and Kohnen (2004).

3 The question of the degree of orality of sermons is discussed in Kohnen (2004 and 2007c) and has also been discussed in Volk (1988).

4 See also Kohnen (this volume) who discusses such an approach for the genre of prayers.

5 In using the term 'discourse marker', I am following Fludernik's (2000) usage of the term, who defines discourse markers as cohesive means demarcating episode boundaries (in narratives). Discourse markers are not only adverbs and adverbial phrases but also changes in tense, for instance. For the importance of tense as means of cohesion in texts see also Brown and Yule (1983) or Quirk et al. (1985).

6 Note that the fourteenth century is only represented from 1350 onwards, since it is difficult to find extant texts from the early fourteenth century. Most of the extant sermon material, for example, is from the latter half of the century (see Heffernan and Horner 2005) and there is convincing evidence that sociohistorical factors are responsible for this uneven distribution. Heffernan, for example, sees the Black Death as one factor influencing sermon production (at least in their fixation in writing) in the fourteenth century (cf. Heffernan 1984: 195).

7 John Knox was a Scottish reformer. Dialectal features, however, are mostly confined to the levels of lexis and orthography, and the text largely conforms to genuinely English treatises as concerns its structure and function. Knox's biography gives ample evidence that he was widely travelled (ministries included Geneva and Frankfurt). He was also involved in the conception of the *Book of Common Praye*r under Edward VI and many of his works were indeed intended for a wider audience outside Scotland. Knox's *Exposition upon Matthew*, included in this study, was first printed from a manuscript in the possession of the widow of Edward Dering, Puritan reformer and Fellow at Christ's College, Cambridge, and was first printed in London. His *Declaration* (see example (10)) was also published in London even though it bears the fictitious imprint "Imprinted in Rome" (cf. Laing 1855/1966).

8 For the concept of face and for politeness (strategies) see Brown and Levinson (1987).

9 Correlation tests (*Pearson product-moment correlation*) showed a negative correlation of r=-1.

10 Scholastic writing may be considered 'argumentative' in a way and some predecessors of 'argumentation' may be found in phrases in which the author seems to anticipate objections and opposing views from his audience and reacts as if they had actually been voiced: "Parauntur *it is replied aȝeyns me þat* þer be many holy men, al-be-it þei com not to þe holynes of oure old holy faders. *I answere* as Seynt Barnard seith," (anon., *Middle English Sermons*, 1400-1415). However, neither of these constitutes a distinct pattern. A similar claim was already made by Matti Rissanen in 1996 when discussing the *Helsinki Corpus*: Rissanen says that argumentation is a relative latecomer in the history of English, but that argumentative passages may be found in instructive writing on a small scale (1996: 233).

11 Both Knox (a reformer) and Fisher (an orthodox catholic) are quite similar in how they utilize exhortation and it remains to be seen if different denominations differ in any systematic way in their usage of exhortation. The samples of reformists like Knox, Latimer and Ridley and those of people like Fisher writing against the Reformation, included in this study, do not show any significant differences in this respect.

12 O'Mara (1994, 2002) has pointed out the importance of "single or small group" sermons (1994: 1) against the overwhelming majority of sermon collections and cycle sermons, particularly in the fifteenth century (see also Heffernan 1984).

References

Arnovick, L.K. (1999), *Diachronic Pragmatics. Seven Case Studies in English Illocutionary Development*. Pragmatics & Beyond New Series 68. Amsterdam: John Benjamins.

Barratt, A. (1984), 'Works of religious instruction', in: A.S.G. Edwards (ed.) *Middle English Prose. A Critical Guide to Major Authors and Genres*. New Jersey: Rutgers University Press. 413-432.

Bell, A. (1984), 'Language style as audience design', *Language in Society*, 12: 145-204.

Biber, D. and E. Finegan (1992), 'The linguistic evolution of five written and two speech-based English genres from the 17th to the 20th centuries', in: M. Rissanen, O. Ihalainen, T. Nevalainen and I. Taavitsainen (eds.) *History of Englishes*. Berlin: Mouton de Gruyter. 687-704.

Brown, G. and G. Yule (1983), *Discourse Analysis*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Brown, P. and S.C. Levinson (1987), *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Day, V. (1974), 'The influence of the catechetical narratio on Old English and some other medieval literature', *Anglo-Saxon England*, 3: 51-61.

Dent, A. (1601), *The Plain Man's Pathway to Heaven*. London: Dexter.

Fludernik, M. (2000), 'Discourse markers in Malory's *Morte D'Arthur*', *Journal of Historical Pragmatics*, 1(2): 231-262.

Green, I. (1996), *The Christian's ABC. Catechisms and Catechizing in England c1530-1740*. Oxford: Clarendon.

Green, I. (2000), *Print and Protestantism in Early Modern England*. Oxford: Oxford University Press.

Heffernan, T.J. (1984), 'Sermon literature', in: A.S.G. Edwards (ed.) *Middle English Prose. A Critical Guide to Major Authors and Genres*. New Jersey: Rutgers University Press. 177-208.

Heffernan, T.J. and P.J. Horner (2005), 'Sermons and Homilies', in: P.G. Beidler (ed.) *A Manual of the Writings in Middle English 1050-1500*, volume XI. New Haven: Connecticut Academy of Arts and Sciences. 3969-4167.

Jucker, A.H. (2008), 'Politeness in the history of English', in: R. Dury, M. Gotti and M. Dossena (eds.) *Selected Papers from the Fourteenth International Conference on English Historical Linguistics (ICEHL 14), Bergamo, 21-25 August 2006*. Volume II: Lexical and Semantic Change. Amsterdam: John Benjamins. 3-29.

Kahrl, S. (1966), 'The medieval origins of sixteenth-century English Jest-Books', *Studies in the Renaissance*, 13: 166-183.

Kohnen, T. (2004), *Text. Textsorte. Sprachgeschichte. Englische Partizipial- und Gerundialkonstruktionen 1100-1700*. Tübingen: Niemeyer.

Kohnen, T. (2007a), 'From Helsinki through the centuries: the design and development of English diachronic corpora', in: P. Pahta, I. Taavitsainen, T. Nevalainen and J. Tyrkkö (eds.) *Towards Multimedia in Corpus Studies*. Helsinki: Research Unit for Variation, Contacts and Change in English. (Studies in Language Variation, Contacts and Change in English, volume 2). Online publication www.helsinki.fi/varieng/journal/index.html.

Kohnen, T. (2007b), 'Text types and the methodology of diachronic speech-act analysis', in: S.M. Fitzmaurice and I. Taavitsainen (eds.) *Methods in Historical Pragmatics*. Berlin: Mouton de Gruyter. 139-166.

Kohnen, T. (2007c), 'Connective profiles in the history of English texts: aspects of orality and literacy', in: U. Lenker and A. Meurman-Solin (eds.) *Clausal Connection in the History of English*. Amsterdam: John Benjamins. 289-308.

Kytö, M. (1996), *Manual to the Diachronic Part of the Helsinki Corpus of English Texts*. 3rd ed. Helsinki: University of Helsinki.

Laing, D. (1855), *The Works of John Knox*. 6 Volumes, Reprinted 1966. Edinburgh: Bannatayne Club.

O'Mara, V.M. (ed.) (1994), *A Study and Edition of Selected Middle English Sermons*. Leeds Texts and Monographs, New Series 13. Leeds: University of Leeds.

O'Mara, V.M. (ed.) (2002), *Four Middle English Sermons*. Heidelberg: Winter.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar Of The English Language*. London: Longman.

Rissanen, M. (1996), 'Genres, texts and corpora in the study of Medieval English', in: J. Klein and D. Vanderbeke (eds.) *Proceedings of the Conference of the German Association of University Teachers of English (Anglistentag), Greifswald 1995*. Tübingen: Niemeyer. 229-242.

Sargent, M.G. (1984), 'Minor devotional writings', in: A.S.G. Edwards (ed.) *Middle English Prose. A Critical Guide to Major Authors and Genres*. New Jersey: Rutgers University Press. 147-176.

Taavitsainen, I. and P. Pahta (1995), 'Scientific thought-styles in discourse structures: changing patterns in a historical perspective', in: B. Wårvik, S-K. Tanskanen and R. Hiltunen (eds.) *Organization in Discourse: Proceedings from the Turku Conference. Anglicana Turkuensia 14*. University of Turku. 519-529.

Taavitsainen, I. and A.H. Jucker (eds.) (2003), *Diachronic Perspectives on Address Term Systems*. Pragmatics & Beyond New Series 107. Amsterdam: John Benjamins.

Volk, S. (1988), 'Pulpit rhetoric and pastoral care. An approach to problems of literacy and orality in late fourteenth and early fifteenth century vernacular sermons in England', in: W. Erzgräber and S. Volk (eds.) *Mündlichkeit und Schriftlichkeit im englischen Mittelalter*. Tübingen: Narr. 147-163.

Walker, T. (2003), 'You and thou in Early Modern English dialogues – patterns of usage', in: I. Taavitsainen and A.H. Jucker (eds.) *Diachronic Perspectives on Address Term Systems*. Pragmatics & Beyond New Series 107. Amsterdam: John Benjamins. 309-342.

Werlich, E. (1976), *A Text Grammar of English*. Heidelberg: Quelle and Meyer.

Appendix

This corpus is based on the *Helsinki Corpus* (cf. Kytö 1996) and has been supplemented with texts from the *Corpus of English Religious Prose* (currently being compiled at the University of Cologne, see www.helsinki.fi/varieng/CORD /corpora/COERP/index.html).

	Author	Text	Size (no. words)
Sermons			
14th c	Hugo Legat	In Passione Domini	5,128
	Anon.	Lenten Sermons	5,179
15th c	John Mirk	Festial	5,029
	Anon.	Middle English Sermons	5,044
	Anon.	Single Sermons	4,670
	Richard Fitzjames	Sermo Die Lune	5,038
16th c	John Fisher	Sermon	4,853
	Hugh Latimer	Sermon on the Ploughers/Seven Sermons	5,066
	Anon.	Elizabethan Homilies	5,065
	Henry Smith	Sermon "of usurie"	5,025
17th c	Richard Hooker	Sermon upon St. Jude	5,005
	Anon.	Sermons	5,013
	John Tillotson	Scoffing at Religion	5,082
	Jeremy Taylor	The Marriage Ring	5,055
Treatises			
14th c	Anon.	The Lay Folk's Catechism	5,027
15th c	Richard Ermyte	Exposition of the Lord's Prayer	5,001
	Anon.	Vices & Virtues	5,003
	Richard Lavynham	Litil Tretys	5,024
	Anon.	Treatise of Love	5,006
16th c	John Knox	An Exposition	5,009
	John Knox	A Declaration	5,045
	Nicholas Ridley	A Pituous Lamentation	4,995
	Francis Bacon	Religious Meditations	4,373
17th c	John Donne	Essays in Divinity	5,079
	Joseph Hall	Contemplations	5,063
	William Penn	The Christian Quaker	5,076
	John Bunyan	The Barren Fig-Tree	5,063

Friends will be "friends"? The sociopragmatics of referential terms in early English letters[1]

Minna Nevala

University of Helsinki

Abstract

This article studies the sociopragmatic use of the term friend *in personal correspondence in Early Modern and Late Modern English, more specifically, during the seventeenth and eighteenth centuries. The analysis shows that* friend *may be used, firstly, for an instrumental function, as when the writer has something to gain from it: a favour, a reciprocal act of solidarity, or access to the addressee's/referent's in-group. Secondly, writers use* friend *in emotional contexts, for example when expressing intimacy and affect towards the referent. The material also includes other functions for the term, such as in expressions of goodwill or in news reporting. In general, shifting between in-group/out-group membership appears to be a common function for the use of* friend*. Over time, the term is increasingly used to indicate intimate friends and acquaintances, whereas reference to members of the family and kin as "friends" declines from the latter half of the seventeenth century onwards.*

1. Introduction

Speakers often use referential terms to express their relationship not only to the referent but also to the addressee. With the help of such terms, which usually encode both the speaker-referent and the speaker-addressee relations, the hearer or the addressee is able to calculate the actual or inferred distance between him/herself and the speaker (cf. Murphy 1988). Referential expressions may thus be used to convey the speaker's power over the addressee and/or the referent, as well as to alter distance between the interactants.

Previous studies (Nevala 2004a, 2004b and forthcoming) show that using person-referential terms may be seen as a device both for claiming the writer's own status as authoritative enough to reduce the distance between the interactants, and for boosting the recipient's or the referent's importance as a prospective or existing member of the writer's in-group. For example, before the eighteenth century, the term *friend* was most often used in direct address between social equals, and it functioned both as an indicator of intimacy and as an expression of solidarity. Later in the eighteenth century the reasons for calling someone a "friend" changed somewhat: it appears that the term *friend* became a more strategic tool for promoting the image of, if not the other, then oneself.

In this article, I will study the sociopragmatic use of the term *friend* in third-person reference in the seventeenth and the eighteenth centuries. My material comes from the *Corpus of Early English Correspondence* (CEEC), its *Supplement* (CEECSU), and its *Extension* (CEECE), which consist of personal letters from c. 1400-1800. My purpose is to study, firstly, who those referred to as friends are, i.e. what their social distance is relative to the writer. I aim to test whether my material indicates the gradual semantic change in the meaning of the term from 'a member of the family and kin' to 'an intimate friend', which sociohistorians like Stone (1977) and Macfarlane (1986) claim happened at around the turn of the eighteenth century. Secondly, I will take a look at how the discourse functions underlying the choice of referential terms have changed over time, concentrating on the instrumental and the emotional uses of *friend* in particular.

2. Social deixis and referential terms

My approach in the study of person reference is based on the hypothesis that the use of referential terms is deeply rooted not only in social hierarchy but also in individual social roles. In order to describe how referential terms are used in historical material, one has to go beyond the immediate situational context into the social, as well as societal, dimension.[2] The form which language takes in my material can be said to be influenced, and in some cases even dictated, by social conventions. Historical sociopragmatics is used here as a term to cover not only the diachronic variation and change at the social and societal, i.e. in Leech's (1983) terms "local", level of language use, but also in the individual and contextual dimensions. In other words, it comprises the macro-level of social, socio-cultural and sociological factors, as well as the micro-level of personal, situational and stylistic factors.

Referential terms are intrinsically deictic expressions. In Levinson's (2004: 97) words, deixis "introduces subjective, attentional, intentional and, of course, context-dependent properties into natural languages". In Sidnell's (2007) view, deixis as a whole is a sociocentric rather than an egocentric phenomenon, which is based on the ability of the participants to make interactional calculations on each other's "perspectives". Social deixis in particular involves the speaker-referent relation, which means that by using, for example, direct address or referential terms the hearer or the addressee may calculate what the relation is between him/herself and the speaker by looking at both the speaker-referent and the speaker-addressee relations (Levinson 1979, 1988, 1992, 2004; Sifianou 1992; also Traugott and Dasher 2005). Referential terms, like direct address terms, may be used to alter distance in respect to the addressee and/or the referent. The speaker may also use reference for shifting power relations by means of shifting between terms marking in-group or out-group characteristics (e.g. possessive pronouns *my*, *our* vs. *your*, *their*).

2.1 Social selves and shared knowledge

In a communicative situation, the interactants need to have a shared background or contextual information in order to be able to understand to what or whom the respective speaker is referring (cf. Schegloff 1996; Hacohen and Schegloff 2006). According to Krauss (1987: 86), the meaning of a message can be thought of as something that is negotiated between the speaker and the addressee (receiver), and this requires a shared base of knowledge. Studies on reference in general agree that mutual knowledge is a central prerequisite in the process of determining the referent (see e.g. Clark and Marshall 1981). Murphy (1988: 340), for example, mentions "mutual information" as the basis of cooperativeness in his Rule of Polite Reference (RPR). Hanks (1992: 59, 67ff.) also writes about the sufficient background knowledge, "a common framework", that the speaker and the addressee must have for the reference to be successful. Other writers talk about "common knowledge" (Evans 1982) and "common ground" (Clark and Murphy 1982). Schegloff (1996: 459) divides the use of background knowledge in reference into "recognitional" and "non-recognitional". When using the former, both the speaker and the addressee know who the referent is; in the latter, only the speaker may be able to identify the referent.

When people communicate with those they do not know well, they can first make only rough assessments of what the others are likely to know. Hogg and Abrams (1988: 195) have found that an individual's speech repertoire in general contains language varieties which fulfil a social function in conveying a framework of shared meanings. The process of finding out what the shared knowledge is between the interactants usually starts with categorising others into social categories: the more that is known about a social category, the more accurately the speaker can predict what kind of knowledge a typical member of that category is likely to have. After that, the speaker will be able to better assess the addressee's informational needs in the course of communication. In a larger perspective, shared knowledge is important in building and indexing 'social selves' which derive from experience in interpersonal relationships and memberships in impersonal collectives and social categories (cf. Brewer and Gardner 1996).

2.2 Person reference as a marker of social group membership

In this sense, referential terms relate to the concepts of in-group convergence and out-group divergence. Referent honorifics, for example, are particularly sensitive to in-group membership. In reference, as well as in direct address, the relationships between the speaker, hearer, referent, overhearer, and bystander, etc., can be said to be influenced not only by social distance, but also by psychological distance or power. The speaker may wish to express whether he is a member of a certain in-group or out-group, in relation to either the recipient or the referent or both, by referring to a third person with a term which can easily be interpreted. Writers who are socially or psychologically superior or equal to their addressee or

referent may wish to stress their status by using intimate terms of reference for someone in their own in-group. Alternatively, a writer who is inferior to the recipient may wish to give the impression that he is e.g. a close member of the referent's in-group.

In his study of audience design, Bell (1984) distinguishes an axis of style which includes what he calls "referees". These are third persons who are not physically present in a communicative situation yet influence the language the speaker (or writer) uses, whereby the speaker shifts from addressee-oriented to referee-oriented style. Referee design can then be further divided into in-group design (identifying with the speaker's own group) and out-group design (oriented towards an outside reference group). Bell links style variation at this level to the accommodation model which hypothesises that a person accommodates his/her speech to the addressee in order to be approved (for the accommodation theory, see also e.g. Giles and Smith 1979; Coupland and Giles 1988).

The use of reference terms may of course be taken as an example of a certain type of referee design. In my previous studies (Nevala 2004a and 2004b) I found that, in letters, the writer may wish to express whether s/he is a member of the in-group or the out-group by referring to a third person with a term which can be easily interpreted. Writers socially superior or equal to their addressee and/or referent may want to stress either their superiority or in-group properties by using intimate terms of reference for someone in their own in-group (but not of the addressee's). Murphy (1988: 328) uses the term "name-dropping" for this kind of flaunting of social advantage against a normal situation in which it could be socially awkward to use, for example, a first name as a reference term if the addressee did not enjoy an equally intimate relationship with the referent as the speaker.[3] On the other hand, there are situations in which a speaker is socially inferior to the addressee and/or the referent. Also in these cases, the speaker may wish to give the impression that s/he is a close member of the referent's in-group, although it is clear from the context that s/he is not. In my earlier study, the writers often emphasised the desired closeness of the relationship by using terms like *my intimate friend* or *my very good friend*.

According to Perdue et al. (1990: 476; see also Hogg and Abrams 1988: 74) terms referring to in-group categorisation, such as the pronoun *we*, may over time accumulate connotations that are primarily positive, whereas out-group-referent words (such as *they*) are more likely to have less favourable, even negative, connotations. Pronouns in general introduce evaluative biases to new and unfamiliar addressees and help establish positive or negative predispositions. Such predispositions breed stereotypes: members of the in-group tend to be seen as favourable and out-group members as unfavourable. Similarly, nominal reference terms may be used for the purposes of inferring positive or negative characteristics: it is more preferable, for example, to be a "friend" than to be an "enemy".

3. The use of *friend* in early English society

My earlier study (Nevala 2004a) shows that the term *friend* was often used in direct address between social equals, most commonly between members of the lower gentry, the lower clergy and professionals. The term might imply true friendship and intimacy, which was usually the case with gentry correspondents and, before the eighteenth century in particular, often with relatives by blood or marriage, or it could be used as an expression of solidarity, as occurred mostly in the early letters of the clergy. If *friend* was used by a superior in power to an inferior, it was most likely to appear in letters from a husband to a wife or, if the social distance was greater, in cases where the relationship between the writer and the addressee was somehow closer than that between average strangers (e.g. when the correspondents were in frequent contact with each other), or when the superior wanted to express goodwill to the addressee. By contrast, when *friend* was used by an inferior of a superior, it might work as a kind of social 'softener', a booster of the recipient's or the referent's authority or, in direct address, a device to assure the addressee of the writer's loyalty.

In his study of marriage in historical England, Macfarlane (1986: 146-147; see also O'Hara 1991) found that the "friends" of a married couple might include their employers, their guarantors, guardians or more distant relatives. In general, obtaining the "goodwill of friends" was important to couples intending to marry: it was wise to obtain the support of those who might help with connections, advice, and gifts.[4] Tadmor (2001: 167) agrees that in the seventeenth and eighteenth centuries, *friend* was used to refer not only to kinship but also to a wide range of non-related supporters, such as patrons, employers, allies, well-wishers, companions, members of social circles, and intimate friends. The term had several meanings that ranged from kinship ties, emotional relationships, economic and occupational connections, intellectual and spiritual attachments to social networks and political alliances. Stone (1977: 79) in turn makes a division between sixteenth- and seventeenth-century "friends", arguing for different meanings of the singular and plural uses of the term. *Friend* in the singular often meant a loved one, but also "someone who could help one on in life, with whom one could safely do business, or upon whom one was in some way dependent". The plural meaning of the term was used to refer to advisors, associates, and supporters of any kind, as well as to relatives by blood or marriage, neighbours, or people of high social status where there was hope of patronage.

The eighteenth-century situation is somewhat different from the earlier centuries. Sociohistorians, such as Macfarlane (1986: 145), claim that whereas in the earlier periods the word *friend* had been used for relatives as well as non-relatives, in the eighteenth century it started to be used mainly in its modern, more emotionally-laden, meaning. For example, in his dictionary Dr. Johnson defined *friend* as "someone who supports you and comforts you while others do not", and "someone with whom to compare minds and cherish private virtues" (Stone 1977: 79). In a later edition, a *friend* meant "an intimate, a confidant, a favourer".[5]

It is my purpose in this article to compare whether diachronic variation in the meaning and use of *friend* in the course of the seventeenth and eighteenth centuries does indeed occur. If the sociohistorians' hypotheses are correct, the authentic letter material I have used for my analysis should indicate a clear increase of the emotional discourse function over the instrumental function of the term within the period. The use of *friend* in reference to members of the nuclear family and other kin should also show a clear decrease from the seventeenth to the eighteenth century, whereas references to intimate friends should increase in number.

4. Material and method

The material for this study comes from the *Corpora of Early English Correspondence* (CEEC400), including the CEEC proper, its *Supplement* and eighteenth-century *Extension*. The corpora in question cover almost 400 years of personal letters, and together add up to c. 5.2 million running words. The figures in table 1 include all instances of *friend* as a third-person referential term in the seventeenth- and eighteenth-century part of the corpus material, which add up to a total of 3,841 instances. The figures show that there is a clear increase in the number of instances from the seventeenth to the eighteenth century (from 9.03 to 10.86 when normalised per 10,000 words).

Table 1: The material

Time period	Running words	Instances of *friend(s)*	Instances /10,000 words
17th century	1,931,278	1,743	9.03
18th century	1,932,683	2,098	10.86
Total	3,863,961	3,841	9.94

The figures do not include instances of *friend* in proverbs and sayings, such as "God to friend", or in its verbal use, such as "to make friends", "to friend", and "to befriend". Similarly, I left out all those cases in which *friend* is used in its generic sense (e.g. "everyone needs friends"), or those of direct address, including letter superscription formulae. I have, however, counted those cases in formulaic sequences in which the writer is securing goodwill or sending regards to many identifiable referents, often including also the recipient. Otherwise, instances which refer to the recipient or the writer him/herself have not been included in the figures.

5. Distance between "friends"

In order to study both synchronic and diachronic variation of interpersonal distance in the use of *friend*, I have divided the instances into four separate categories according to the relationship between the writer and the referent: 'family', 'intimate friends', 'acquaintances', and 'more distant' people (including total strangers). The division has been made on the basis of sociohistorical facts, so that the 'family' category comprises the members of the nuclear family and other kinship; 'intimate friends' include people to whom the writer has not only close contact but also personal affect; 'acquaintances' are referents in frequent contact with the writer, but also favourers, patrons, and general supporters; and the 'more distant' category includes e.g. friends and acquaintances of the recipient or of other people, supporters of political views/parties favoured by the writer, or the so-called "friends of the state". The difference between 'intimate friends' and 'acquaintances' has not been an easy one to make, since often the immediate context does not offer enough clues for deciding whether the relationship functions more on the basis of personal trust and affection than of frequency of contact.[6] Any lack of contextual clues has mostly been complemented by checking background information on the writer/referent or the recipient/referent relationship from the original letter editions.

Table 2: Social distance between the writer and the referent identified as a "friend"

Writer-referent relationship	17th century	/10,000 words	18th century	/10,000 words	Total	/10,000 words
Family	345	1.79	123	0.64	468	1.21
Intimate friends	127	0.66	407	2.11	534	1.38
Acquaintances	745	3.86	911	4.71	1,656	4.29
More distant	526	2.72	657	3.40	1,183	3.06
Total	1,743	9.03	2,098	10.86	3,841	9.94

Table 2 includes all instances from the material in both centuries and their relative distribution per 10,000 words. As seen in table 1, the total number of instances increases from the seventeenth century to the next. Correspondingly, the use of *friend* rises in all relationship categories except one: the figures in the 'family' category decline from 345 instances to 123 (1.79 to 0.64 per 10,000 words). If we take a closer look at the distribution of *friend* within the four categories, we can see that the term is mostly used to refer to "acquaintances" throughout the material (1,656 instances; 4.29 per 10,000 words). "Intimate friends" is the smallest reference group in the seventeenth century, whereas in the eighteenth century the meaning which is least inferred by the term is "family". Also in the total number of instances, the 'family' category remains the smallest with 468 examples (1.21 instances per 10,000 words). All figures in table 2 are statistically highly significant ($\chi^2 = 253$, df = 3, $p < 0.001$).

Examples (1) to (5) show typical representations of each category from the material. In (1), the writer refers to his family members, and possibly to other kin, with "loving frendes". Positive attributes are also used to modify *friend* in example (2), in which Samuel Crisp writes about his intimate friend Molly Chute.

(1) Most deare & loving mother, my most humble duty remembred with the frendly remembrance of all my sisters & loving frendes in generall. My last l'res unto you was by a shipp called the Hoziander, w'ch went no farther then Bantam & stayed in these p'tes, so that I make doubt of conveyance of my l'ers unto you. (FACTOR3: Richard Wickham, 1617)

(2) Molly Chute (an intimate & most infinitely agreable [sic] old friend of mine, long since dead) when I us'd to desire her to love me a great deal, would say *Look Ye Sam, I have this Stock of love by me*, putting out her little Finger, *& I can afford You so much*, measuring off perhaps half the length of her Nail - *& I think that's'* [sic] *pretty fair* (BURNEYF: Samuel Crisp, 1778?)

Example (3) includes reference to acquaintances, "our country friends", with whom the writer is in (frequent) contact, but with whom she has a more casual relationship compared to the seemingly affectionate friendship portrayed in the previous excerpt. Example (4) in turn shows a contemporary view on the difference between "friends" and "acquaintances".

(3) I hear the fitters have sign'd so that point I hope will be easy to you all for the future. I've order £40 of the mony you've return'd to be paid Mrs. Caverley, and when our country friends come we shall know what more they've occasion for. (CLAVERING: Anne Clavering, 1709)

(4) And tho' I am oblig'd to you for your polite professions – yet your benevolence seems divided amongst such a numerous acquaintance – that a very small share of it can be reserv'd for those – whom you [absurdly call] honor with the title of your *Friends* (DODSLEY: Richard Graves, 1759)

(5) A sett for M^{r} Jago I have sent by his direction to a relation of his in town, who will forward it to him. And if you have any other friends that you would chuse to make presents to, I beg you will without the least scrupple send me your com~ands. (DODSLEY: Robert Dodsley, 1755)

An example of the use of *friend* to imply a more distant relationship is shown in (5), in which the writer refers to the friends/acquaintances of the recipient with "any other friends". As mentioned above, instances in this category show the greatest variation between different referent statuses relative to the writer (e.g. a friend of the recipient or a supporter of a political party like the Whigs).

6. Discourse functions of *friend*

The instances of *friend* found in the material have been divided into three main categories according to their functions in the immediate context. The category of instrumental friendship includes those cases in which either being or having a friend benefits the party in question in some way, mostly by material gain, be it money, a favour or help, a compliment, employment, protection or crucial information. Emotional friendship mainly includes those cases of mental support in which the person, be it the writer, the recipient or the referent, expresses his/her personal affinity with the person in question, or shares his/her feelings about the person. There are also other functions which include the instances of *friend* in e.g. formulaic use and news reporting. Also, cases in which more than one discourse function is simultaneously employed are placed in the 'other' category.

Table 3: Discourse functions of *friend* in the material

Function	17th century	/10,000 words	18th century	/10,000 words	Total	/10,000 words
Instrumental	761	3.94	800	4.14	1,561	4.04
Emotional	222	1.15	358	1.85	580	1.50
Other	760	3.94	940	4.86	1,700	4.40
Total	1,743	9.03	2,098	10.86	3,841	9.94

Table 3 contains the distribution of *friend* as categorised by discourse functions. Diachronically, the use of the term increases in all functions from the seventeenth to the eighteenth century: the categories 'instrumental' and 'other' comprise most cases of *friend* throughout the period (1,561; 4.04 per 10,000 words and 1,700; 4.40 per 10,000 words respectively), while the term remains less used in its 'emotional' function (580; 1.50 per 10,000 words). The difference in the number of instances in the 'instrumental' and 'other' categories grows, however, in the eighteenth century, and the figure for the 'emotional' function increases relatively more than that for the 'instrumental' function over the course of time. All figures in table 3 are statistically highly significant ($\chi^2 = 19$, cf = 2, $p < 0.001$).

6.1 Instrumental function

In this and the following subsections, I will present examples from each category in more depth. I will begin by discussing instances of the instrumental discourse function, and then move on to the emotional and other functions of *friend*.

In the first example from the seventeenth century, (6), Arabella Stuart describes how she is thankful for the favour from her "very honourable frend", and how the said favour has helped her into the distinguished circle of the referent's friends, "such great persons". The term "his great ffriend" in example (7) also concerns instrumental gain, or rather, the lack of the proper kind of "Service" to the Bishop in question, except for mere news reporting. Example (8),

on the other hand, concerns promises of gain yet to be fulfilled by "friends at court", which, as the writer sarcastically notes, would probably never amount to anything unless money was to first change hands.

(6) I finde him my very honorable frende, both in word and deede, I pray you give him such thancks for me as he many wayes deserves. and especially for this extraordinary and unexpected favour, whearby I perceive his Lordship reckneth me in the nomber of his frends for whom onely such great persons as he reserve such favours. (STUART: Arabella Stuart, 1604)

(7) He was an acquaintance of ye Bishops, and had flatter'd him into an opinion of being his great ffriend, tho' I could never learn, either by my own knowledge whilst I lived with ye B~p; or by any enquiry I could make, that he had ever done him any Service, further than to oblige him sometimes with a letter of News or Occurrences from London. (FLEMING: Henry Brougham, 1694)

(8) *Mr. Ralfe Smith* the Carver hath an abundance of *friends at court* amongst the Lords and others, and hee hath fair shews and promises enough from them all *even from the King himselfe* but *my brother Richard saies that* without mony hee will never get anything *but* promises amongst them; and if hee cannot I know not who can. (OXINDENX: Henry Oxinden, 1663?)

The instrumental use of *friend* in the eighteenth century is rather similar to its use in the previous century. In example (9), the addressee is in a way included in the general category of a "friend", as the writer mentions how the recipient, rather "than a Landlord", deserves the gift of a Christmas turkey which is the only present in the writer's power to give. In example (10), the gift is a recommendation to one "friend Hey", who, according to the writer, is "in high reputation at Cambridge", and thus able to help the recipient's son, "this unhappy boy", to procure a student position at the university. Later correspondence shows, however, that both Twining's and Hey's efforts were fruitless: another boy was chosen for the position instead.

(9) a goose, or Turky is carry'd with great form to the Squire, and tho' he wants them not, & has much better of his own, yet he never fails to accept them, and, if he happens to be, what they call, a well-spoken man, with expressions of great civility and respect. Permit me, Sir, to imitate this laudable custom, and 'tho' there is no want of amusements at Norton, please to accept this, as my Xstmas Offering. A friend, I think, has a greater right to them, than a Landlord; and as this is the only one in my power to make, let not the meaness of it be any Objection. (HURD: Richard Hurd, 1742)

(10) Tho' I never saw this unhappy boy, the manner in which I had heard you speak of him had added to the interest which his name alone wou'd have given me for him. I know nobody of his college; all I cou'd do was to recommend him to my friend Hey, who is Tutor at Sidney College, & in high

reputation at Cambridge, both for his character & abilities. (TWINING: Thomas Twining, 1777)

(11) A Gentleman at Cambridge, an intimate Friend of my Brother's, has undertaken to manage all my little Concerns in that place. When the Books and Furniture are sold, I shall remit the Money to you, and then be obliged to you for your Help in disposing of it to the best Advantage. (COWPERW: William Cowper, 1770)

Another example from the eighteenth century, (11), deals with a helpful "friend" in Cambridge. William Cowper refers to a gentleman, "an intimate Friend" of his brother's, who has helped him to sell his belongings in order to send money to the recipient.

6.2 Emotional function

The emotional use of *friend* in the seventeenth-century material often concerns condolences on the loss of a dear one or empathy for the referent's misfortune. In example (12), Charles II writes how "sensible" he is of the sufferings of the recipient's friends, while Phillip Henry goes to more extreme measures in (13) when expressing his "hearty Sympathy", "weeping with those that weep", and thus identifying with the grief of the "Afflicted friends".

(12) Pray lett all your frinds know how sensible I am of their sufferings, knowing it is only for my sake, and that I am very much grived that I am not in a better condition to lett them see it, but I hope myne will mend, and then I am sure there's shall be better. (HAMILTON: Charles II, 1650)

(13) Wee have hearty Sympathy with our Afflicted Friends in their Afflictions, weeping with those that weep, as being our selves also in the Body, Natural, Mystical. (HENRY: Phillip Henry, 1685)

(14) I frequently drink Your Health with Lord Harley who is always the same good Man, and grows dayly more beloved & more universally known[.] I do so too with our honest and good natured friend Ford, whom I love for many good reasons and particularly for that He loves You. (SWIFT: Mary Prior, 1721)

(15) Now then, my brother (who I begin by saying is the best freind I have in ye. world, & to whom I am more attach'd then to anything else upon ye. face of the earth excepting yourself) has always been my father's favorite & always has been treated as such, (tho' yt. never made ye least alteration in my brother's conduct to me) I understand yt. (GEORGE4: George IV, 1785)

More positive examples of the emotional function of friendship can be found in the eighteenth-century data. The greater emotional involvement, or rather the more verbally prolific expression of it, appears to continue and develop during this century. Mary Prior's straightforward declaration of love towards the referent, and at the same time towards the recipient, in (14) is a good example of

the style which is characteristic of many eighteenth-century letter-writers. George IV's reference to his brother in example (15) is no exception either; a number of writers use superlatives when describing their "dear friends" and family members.

6.3 Other discourse functions

The formula of sending greetings or commendations to the family, friends, and neighbours of the recipient appears most common in the early part of the material and slowly decreases towards the end of the eighteenth century. The structure of the formula in letters to family members and kin remains the same throughout the period: first nearest kin (spouse, parents, children), then other relatives, and finally friends and acquaintances (example (16)). In letters from the eighteenth century, however, the contents often show more variation, as is shown by example (17).

(16) All thinges have gon so crosse wth me as I can nether thinke or write as I would. Comend me to your father and all his familie And all other my frendes thereaboutes And so god keepe you. (STOCKWELL: Anthony Antonie, 1604)

(17) Remember me to those friends that do me the honour to enquire after me. I am, Dear S^{r}, your Sincere Friend & humble Servant John Gilbert-Cooper/ Jun^{r}. (DODSLEY: John Gilbert-Cooper, 1747)

Friend also often appears in the context of (news) reporting or gossiping. This kind of reference to "friends" can mean a general listing of events, as the death of a friend's brother in example (18), or it can involve more specific descriptions, like the way in which Charles Burney writes about Jimmy Mathias's "singing" in example (19).

(18) Your frend Christians brother hath buried his fayre younge wife: and the Lord Delaware is lately dead and some say the Lord Stafford. (CHAMBERLAIN: John Chamberlain, 1602)

(19) *á propos* - here has been our Frd Jemmy Mathias to day, singing like a bird - of wisdom, as he is - & has taken up all the precious time I intended to bestow on you, with his old songs & saws – Honour & Aaa Aa Aaa AaArms - &c – (BURNEY: Charles Burney, 1779)

Finally, there are some other types of contexts in which *friend* is used in the material. In example (20), Philip Gawdy refers to "friends" in relation to the public matters of the state, or in the meaning 'favourers or citizens of a particular country'. The use of *friend* in this kind of context is common throughout the material, and in connection with favouring political ideologies in particular, as can be seen in example (21), in which Daniel Defoe sarcastically refers to the Whig party members as "Our old Friends".

(20) the Hollanders parte had muche the better, wher as the Hollande Englishe hauing a great advauntage of the other Englishe did forbeare, and tolde the cheife governer that they colde not fynde in their hartes to massacre their countrymen and frendes in that dystress, wherevppon the cheife commaunder swore and sent them awaye, and sent certayne Wallons and Ducthe whiche cutt all their throotes, when captayne Aderton, and some two or three captaynes, and lyftenantes more wer slayne. (GAWDY2: Philip Gawdy, 1605)

(21) I have had but Little Time Since My Return to look among Our old Friends The whigs, and Therefore Could Say but Little when you were pleased to ask me of Them. I am Sorry to be Witness to So Much of The weakness of Those I Thought would have before Now have been Wiser. (DEFOE: Daniel Defoe, 1711)

7. Possessive pronoun + *friend* as a marker of group membership

The notions of power, status and authority are closely related to distance and deference. In reference, as well as in direct address, the relationships between the speaker, hearer, and referent can be said to be influenced not only by social distance, but also by psychological distance. As already mentioned in section 2, the writer may wish to express whether he is a member of a certain in-group or out-group, in relation to either the recipient or the referent or both, by altering e.g. pronouns, such as *my*, *our*, *your*, *their*. The sociodeictic qualities of possessive pronouns alone facilitate the indexing of in-group/out-group variation, and when used to modify the term *friend*, the effect is further strengthened.

Numerous instances of the possessive pronoun + *friend* combination as a marker of group membership can be found in the material, and a few of them are presented here to exemplify variation in interpersonal distance. Example (22) shows an instance where Erasmus Darwin refers to his close friends Matthew Boulton and James Watt with the term "my mechanical friends". Darwin uses these kinds of terms often instead of names to express his close relationship with his intimate friends. Here he also relies on shared knowledge: Darwin expects the recipient to know to whom he is referring as well as to recognise his referent status in the out-group from Darwin's immediate friends.

(22) My so long neglecting to send you the machine, I promised, can only be excused by my desire of making it better worth your attention - it remained many months at Birmingham before my mechanical friends return'd from fire-engine building in Cornwall. (DARWIN: Erasmus Darwin, 1779)

Example (23) in turn comprises several pronominal shifts within a short stretch of writing. Lucy Russell first refers to the recipient's cousin and neighbour, and then changes into the formulaic "my Lady" when reporting news of her own acquain-

tances. But when she refers to Lord Mounteagle, she switches to the pronoun *our* in "our noble friend" and, by doing so, includes the recipient in the in-group.

(23) Your cousin Killegrew is gonne to see your neyghbour for a while, nothing altered. My La. Uvedale is become the fonde mother of a sonne. My La. Marquis of Winchester is dead, and our noble freind my Lord Mounteagle very ill of a swelling in his throat. John Elviston died on Tuesday last, to the great griefe of all good daunsers. (CORNWALLIS: Lucy Russell, 1614)

There are also examples in which the writer uses the pronoun *your* when referring to the recipient's friends, but at the same time includes him/herself in the recipient's in-group, into the society of his/her friends. In (24), William Pulteney writes about how the recipient's possible arrival to meet his friends and acquaintances would be appreciated by all of them. By the use of the pronoun "myself" and the phrase "others of your friends", as well as the following inclusive pronoun "us", the writer manages to convey in-group membership.

(24) Lord Bolingbrook, Ld Bathurst, Pope, my self & others of your Friends, are got togather in a Country neighbourhood, which would be much enliven'd if you would come and live among us. (SWIFT: William Pulteney, 1731)

(25) Your friends that tell you this will not cure you full little know what your case had been by this time if you had not taken this course you have done. A decumbency of sharp humours might easily have bereft you of the use of your limbs long before this in such a case as yours, which is not properly the gout, as they imagine. (SYMCOTT: John Symcotts Jr, 1633)

Example (25) from John Symcott Jr.'s letter concerns medical advice given by the recipient's own friends. The writer's disbelief in both the quality of the advice and the learnedness of the advisers comes through in the passage, which ends with "as they imagine". Although Symcott is probably, at least, a distant acquaintance to the recipient's "friends", by using the pronoun *you*, he makes it clear that he wishes to remain in and be recognised as belonging to the out-group.

8. Discussion

8.1 Social friends

It has been my purpose in this article to study those that are referred to as friends: who they actually are, and what their relative distance to the writer is. As we saw in section 5, those who are called friends in the material range from spouses and kin relations to business partners and other supporters. Intimacy most often correlates with emotional friendship, and relative power with instrumental

friendship. There is, however, some individual variation, particularly in the letters of the more educated and/or literary eighteenth-century writers, like Richard Hurd, Ignatius Sancho, Charles and Fanny Burney, and Hester Piozzi. These writers seem to use *friend* more frequently and of a greater variety of people than other writers of their time, and their emotional use of the term also exceeds that of others.

The analysis also shows that the use of *friend* in reference to members of the nuclear family and other kin decreases from the mid-seventeenth century onwards, whereas in other relationship categories the use of the term gradually increases. The rise in the 'intimate friends' category supports the sociohistorians' view of *friend* having been used more often in reference to family and kin before the eighteenth century, and, on the other hand, more in its present-day sense of intimate friendship already in the Late Modern period.

Distinguishing between intimate friends and casual friends, i.e. acquaintances, has not always been an easy task. The distinction has been made not only on the basis of sociohistorical facts, but also by looking both at and beyond the immediate context of a particular instance. Also, it is the immediate context which sometimes gives valuable clues for interpreting to which relationship category *friend* refers; for example, the adjacent occurrence of *friend* and *acquaintance* in phrases like "give my compliments to all my friends and acquaintances" imply that the term *friend* refers to people who are close to the writer. This applies to eighteenth-century examples in particular.

If, on the other hand, we look at the adjectival modifiers usually given to the term *friend*, the situation becomes more complex. Intensified forms of such adjectives as *good* and *loving*, i.e. *very good* and *very loving* can be used of both intimate friends and more casual acquaintances. The differences between the use of simple and intensified forms of a particular adjective seem to correlate more with the discourse functions of *friend* than with its use as an indicator of interpersonal distance. The term *very good friend* is often used in emotional contexts, but even more so in strategic, instrumental contexts in which the writer wants to emphasise, for example, the referent's ability to fulfil his/her needs and wishes.

8.2 Functional friends

In general, the instrumental function in the use of *friend* appears to be more common than the emotional function throughout the material. The sociohistorians' hypothesis of an increase in the emotional discourse function over the instrumental function from the seventeenth to the eighteenth century does not seem to hold, although the increase in the number per 10,000 words is relatively greater for the emotional than the instrumental use of the term (for statistical significance, see section 6). Also, the view presented by historians like Tadmor (2001) that instrumentality and personal affection are closely interrelated concepts does not necessarily apply to my material, since the connection between the use of *friend* in reference to family members/intimate friends and the instrumental function does not appear to be more common than between the term

in reference to family members/intimate friends and other functions. There is, however, a correlation between reference to close friends and the emotional discourse function, and instrumentality is the most typical factor when *friend* refers to casual acquaintances.

There seems to be diachronic change within the functional categories themselves, as certain uses are more common than others from one century to the next. For example, in the 'instrumental' category the use of *friend* in contexts which concern gaining money or property appears to be more common than giving and receiving e.g. information in the early seventeenth-century letters than in the later data. Synchronic variation in the use of *friend* seems to appear in such functional contexts which are emotionally positive rather than negative. It is rather surprising to find that the material contains only a few examples in which the writer expresses negative feelings about the "friend" in question, since even e.g. while lamenting the loss of a "friend", the writer simultaneously expresses his/her love and affection towards the deceased. The only truly negative instances of *friend* mostly appear when the writer denies friendship altogether (i.e. "he is no friend of mine"). The 'other' function category shows the greatest variation of the different contexts in which *friend* occurs. The early part of the seventeenth-century material contains more examples of the formulaic use of *friend* than the latter part and, correspondingly, the number of instances in the context of news reporting appears to be greater in the eighteenth-century data.[7]

In general, the present study corroborates the results from my earlier studies in that, during the eighteenth century, the reasons behind referring to someone as a "friend" changed. As the structure of society was changing and a new middle class forming, it appears that the use of the term *friend* also became a more strategic tool for promoting the image of either the writer him/herself or other interactants in the situation. In this context, it is not surprising to find that what could be called name-dropping increases during this century: writers more frequently indicate the referent's social status and name in particular. Similarly, the instrumental function of *friend* in reference becomes even more strategic when the writer has to gain access to the recipient's and/or the referent's in-group.

Shifting between in-group and out-group membership thus appears to be a common function for the use of *friend*. The additional use of possessive pronouns in the group inclusive and exclusive meanings helps the addressee to weigh and recognise the different kinds of relationships more accurately. Whereas the pronouns *my* and *our* seem to indicate in-group membership, *your* and *their* are usually used to imply that the writer belongs to the referent's out-group. Often writers masterfully juggle between different pronouns in order to give the recipient subtle, or not so subtle, hints about whom they wish to include in their in-group and whom they would prefer to leave in their out-group.

Shared knowledge is a prerequisite to the understanding of not only the referents' identities, but also their social and relational statuses, and the writer usually expects the recipient to know both to whom s/he is referring and what kind of a relationship there exists between him/herself, the recipient and the

referent. Friendship is generally understood in the material as a reciprocal relationship, meaning that at least two of the parties find material gain or mental support in the process, or that is at least the primary aim of calling someone a "friend". The use of *friend* in reference is similar here to its use in direct address, in which it can also function as an expression of both familiarity and solidarity towards the recipient. It is particularly powerful in situations where the writer wants to influence the recipient's attitude towards the contents of the message. Despite obvious similarities between the direct and the referential use of *friend*, a more systematic study is needed in order to be able to make more definite conclusions.

9. Conclusion

In this article I have studied the use of the term *friend* in the seventeenth and eighteenth centuries. It has been my purpose to study the concept of reference in relation to social deixis by drawing conclusions about how the social identities of the interactants and their interpersonal relationships are encoded in the use of referential terms in Early and Late Modern English. In particular, I have discussed how shifts in interpersonal distance can be indexed and managed by using the term *friend*, and how the term can be employed in various discursive contexts, ranging from seeking material gain to news reporting.

Early letter writers had to have experience and shared knowledge of both social and societal norms and constraints, and the choice of referential terms often comprised taking both the recipient and the referent into account. By the end of the eighteenth century the relative status of a "friend" seems to have extended to cover almost every possible person, close or distant. As Tadmor puts it in her study of eighteenth-century shopkeeper Thomas Turner:

> In order to be counted as a 'friend' by Thomas Turner, a person […] could be (1) one of Thomas Turner's relations; (2) his wife; (3) a person with whom he had a close intellectual or devotional affinity; (4) a trusty tradesman with whom he had special business contacts; (5) his tenant or landlord; (6) an officer of the excise. The seventh, and last category of 'friends' […] were the supporters of the Whig interest in the locality, especially the supporters of the Duke of Newcastle. (Tadmor 2001: 174)

When we then consider all the qualities "friendship" appears to have entailed in the Early and Late Modern English period, we can see that the list includes at least such notions as 'gain', 'support', 'duty', 'trust', 'love', and 'devotion'. In reality, however, only a few were able to rise to these great expectations.

Notes

1 The author gratefully acknowledges the financial support received from the Academy of Finland during the writing of this article (project number 114045).

2 The concepts of "social" and "societal" are used by Mey (1993) in his definition of the general dimensions of pragmatics.

3 For one, Ervin-Tripp (1972) has found that any deviation from the form normally used for a person usually signals change in distance, as well as in the overall situation (see also Tieken-Boon van Ostade 1999).

4 In addition to *friend*, especially husbands used the term *fellow* when directly addressing or referring to their wives (Houlbrooke 1984: 102; see also Nevala 2004a for a discussion of the term *be(a)dfellow* and Shiina 2003). Alternatively, parents could call their children their "friends" (O'Day 1994: 90).

5 The earliest occurrence the *Oxford English Dictionary* gives for *friend* in the meaning 'one joined to another in mutual benevolence and intimacy' is in *Beowulf*. Correspondingly, the term is used in the sense 'a kinsman or near relation' even today, although only in the plural ("The prisoner will be handed over to the care of his friends").

6 I have found Spencer-Oatey's (1996) discussion on distance very useful in creating the categories, particularly her definitions on the difference between "friends" and "acquaintances".

7 Since the more fine-grained analysis of the data in each functional subcategory is still in progress, the trends presented here are not yet conclusive (a more detailed analysis will appear in Nevala in progress). The percentual variation in the 'instrumental' category (gaining money or property) is tentatively between c. 40 percent (seventeenth century) and c. 20 percent (eighteenth century), and, for example, the change in the 'other' category (the formulaic use of *friend*) is from c. 60 percent to c. 30 percent from the early to the latter half of the seventeenth century.

References

Bell, A. (1984), 'Language style as audience design', *Language in Society*, 13: 145-204.

Brewer, M.B. and W. Gardner (1996), 'Who is this "we"? Levels of collective identity and self representations', *Journal of Personality and Social Psychology*, 71: 83-93.

CEEC400 = *The Corpora of Early English Correspondence*. Compiled by T. Juvonen, S. Kaislaniemi, J. Keränen, M. Laitinen, M. Nevala, T. Nevalainen, A. Nurmi, M. Palander-Collin, H. Raumolin-Brunberg, A. Sairio, and T. Säily. Research Unit for Variation, Contacts and Change in English, University of Helsinki.

Clark, H.H. and C.R. Marshall (1981), 'Definite descriptions and mutual knowledge', in: A.K. Joshi, B.L. Webber and I.A. Sag (eds.) *Elements of Discourse Understanding*. Cambridge: Cambridge University Press. 10-63.

Clark, H.H. and G.L. Murphy (1982), 'Audience design in meaning and reference', in: J.-F. Le Ny and W. Kintsch (eds.) *Language and Comprehension*, Amsterdam: North-Holland. 287-299.

Coupland, N. and H. Giles (eds.) (1988), *Communicative Accommodation: Recent Developments*. Special issue of *Language and Communication* 8.

Ervin-Tripp, S.M. (1972), 'Sociolinguistic rules of address', in: J.B. Pride and J. Holmes (eds.) *Sociolinguistics: Selected Readings.* Harmondsworth: Penguin. 225-240.

Evans, G. (1982), *The Varieties of Reference*. Oxford: Clarendon Press.

Giles, H. and P.M. Smith (1979), 'Accommodation theory: optimal levels of convergence', in: H. Giles and R. St. Clair (eds.) *Language and Social Psychology*. Oxford: Blackwell. 45-65.

Hacohen, G. and E.A. Schegloff (2006), 'On the preference for minimization in referring to persons: evidence from Hebrew conversation', *Journal of Pragmatics*, 38: 1305-1312.

Hanks, W.F. (1992), 'The indexical ground of deictic reference', in: A. Duranti and C. Goodwin (eds.) *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press. 43-76.

Hogg, M.A. and D. Abrams (1988), *Social Identifications. A Social Psychology of Intergroup Relations and Group Processes*. London: Routledge.

Houlbrooke, R.A. (1984), *The English Family 1450-1700*. London: Longman.

Krauss, R.M. (1987), 'The role of the listener: addressee influences on message formulation', *Journal of Language and Social Psychology*, 6: 81-98.

Leech, G. (1983), *Principles of Pragmatics*. London: Longman.

Levinson, S.C. (1979), 'Pragmatics and social deixis: reclaiming the notion of conventional implicature', *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 5: 206-225.

Levinson, S.C. (1988), 'Putting linguistics on a proper footing: explorations in Goffman's concepts of participation', in: P. Drew and A. Wootton (eds.) *Erving Goffman: Exploring the Interaction Order*. Cambridge: Polity Press. 161-227.

Levinson, S.C. (1992) [1983], *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S.C. (2004), 'Deixis', in: L.R. Horn and G. Ward (eds.) *Handbook of Pragmatics*. Oxford: Blackwell. 97-121.

Macfarlane, A. (1986), *Marriage and Love in England: Modes of Reproduction 1300-1840*. Oxford: Basil Blackwell.

Mey, J. (1993), *Pragmatics. An Introduction*. Oxford: Oxford University Press.
Murphy, G.L. (1988), 'Personal reference in English', *Language in Society*, 17: 317-349.
Nevala, M. (2004a), *Address in Early English Correspondence: Its Forms and Socio-pragmatic Functions*. Helsinki: Société Néophilologique de Helsinki.
Nevala, M. (2004b), 'Accessing politeness axes: forms of address and terms of reference in early English correspondence', *Journal of Pragmatics*, 36: 2125-2160.
Nevala, M. (forthcoming), 'Altering distance and defining authority: person reference in Late Modern English', *Journal of Historical Pragmatics*, 10: 212-233.
Nevala, M. (in progress), '*Friends* and *enemies*: functions of person referential terms in early English correspondence'. To be submitted to *Language, Variation and Change*.
O'Day, R. (1994), *The Family and Family Relationships, 1500-1900: England, France and the United States of America.* Houndmills: Macmillan.
O'Hara, D. (1991), ''Ruled by my friends': aspects of marriage in the diocese of Canterbury, c. 1540-1570', *Continuity and Change*, 6: 9-41.
Oxford English Dictionary. 2nd ed. 1989 (eds. J.A. Simpson and E.S.C. Weiner). Additions 1993-1997 (eds. J.A. Simpson and E.S.C. Weiner; Michael Proffitt), and 3rd ed. (in progress) March 2000- (ed. J.A. Simpson). OED Online. Oxford University Press. <http://dictionary.oed.com>.
Perdue, C.W., J.F. Dovidio, M.B. Gurtman and R.B. Tyler (1990), 'Us and them: social categorization and the process of intergroup bias', *Journal of Personality and Social Psychology*, 59: 475-486.
Schegloff, E.A. (1996), 'Some practices for referring to persons in talk-in-interaction: a partial sketch of a systematics', in: B. Fox (ed.) *Studies in Anaphora*. Amsterdam: John Benjamins. 437-485.
Shiina, M. (2003), 'How spouses used to address each other: a historical pragmatic approach to the use of vocatives in Early Modern English comedies', *Bulletin* (Faculty of Letters, Hosei University), 48: 51-73.
Sidnell, J. (2007), 'Deixis', in: J.-O. Östman and J. Veschueren (eds.) *Handbook of Pragmatics 1998* [Revised version 2005]. Available online at <http://www.benjamins.com/online/hop>.
Sifianou, M. (1992), *Politeness Phenomena in England and Greece.* Oxford: Oxford University Press.
Spencer-Oatey, H. (1996), 'Reconsidering power and distance', *Journal of Pragmatics*, 26: 1-24.
Stone, L. (1977), *The Family, Sex and Marriage in England 1500-1800*. Harmondsworth: Penguin Books.
Tadmor, N. (2001), *Family and Friends in Eighteenth-century England: Household, Kinship, and Patronage*. Cambridge: Cambridge University Press.
Tieken-Boon van Ostade, I. (1999), 'Of formulas and friends: expressions of politeness in John Gay's letters', in: G.A.J. Tops, B. Devriendt and S.

Geukens (eds.) *Thinking English Grammar: to Honour Xavier Dekeyser, Professor Emeritus*. Leuven: Peeters. 99-112.

Traugott, E.C. and R.B. Dasher (2005), *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Self-reference and mental processes in early English personal correspondence: A corpus approach to changing patterns of interaction

Minna Palander-Collin

University of Helsinki

Abstract

*This paper explores linguistic variation and change in the way letter writers position themselves in their letters, the focus being on gentlemen's self-reference (*I*) in sixteenth- and eighteenth-century personal correspondence. Self-reference is understood as a linguistic feature relevant to the identity and interpersonal functions of language. The aim is to identify broad changes in patterns of self-reference, using corpus tools, in the data extracted from the* Corpus of Early English Correspondence *and the* Corpus of Early English Correspondence Extension. *The results indicate that self-reference and self-referential mental processes increased from the sixteenth to the eighteenth century. It is suggested that this development may relate to increasing stance marking and involvement observed in the history of English from 1650 onwards, particularly as self-referential mental expressions often serve interpersonal functions.*

1. Introduction

Personal correspondence is written interaction between letter writers, who negotiate their social positions in the letters they write. Such identity and interpersonal work is carried out through several linguistic means including overt category labels, implicatures, stances, styles, and linguistic structures or systems, but it is impossible to tell which linguistic variables have the capacity to index social meaning without paying attention to the context of use and other co-occurring features (Auer 2007; Bucholtz and Hall 2005). Identity construction in interaction is a complex linguistic and social phenomenon, which is further complicated by the fact that neither linguistic nor social systems remain stable (cf. Kress 1989; Fairclough 1992). Research on language and identity, however, tends to focus on present-day speech communities and rarely addresses long-term diachrony in patterns of interaction, i.e. how expressions of social meaning develop and change. This paper, consequently, sets out to explore linguistic variation and change in the way letter writers position themselves in their letters, focusing on gentlemen's self-reference (*I*) in sixteenth- and eighteenth-century personal correspondence. My aim is to identify broad changes in patterns of self-reference, using corpus tools, in the data extracted from the *Corpus of Early*

English Correspondence (CEEC) and the *Corpus of Early English Correspondence Extension* (CEECE).

Self-reference is one type of linguistic choice where the identity and relational functions of language are evident, as illustrated in examples (1) and (2) (italics added in all examples).

(1) Right honorable and worshipfull father and mother, *in the most lowliest wyse that I can*, *I mekely recomend me unto you*, *desiring* to here of your welfaire and prosperitie, the which *I pray* almyghty Jesu long to continue to his pleasure, and to your most joy, and comforth, and harts ease. (CEEC: Germain Pole to his father-in-law Sir Robert Plumpton, 1499?: p. 139)

(2) To come downe into the Cuntry to do *my duetie* to *my Grandfather*, and yow, *gladly I would*, but *I am unprovided of horse and unfurnished of mony*, if it would pleas yow to *send me my annuitie* which it pleaseth yow to *allow me*, which *I should have receaved* at Midsummer *I could make* the better shifte of 50li which by the whole year yow *give me*, the quarter beeing our Lady day *I receaved* 20li. (CEEC: John Holles to his father Denzell Holles, 1587: p. I, 1)

What we can say in the first person varies according to the situation and the addressee, as the examples above illustrate. Writing to his father-in-law in example (1), Germain Pole refers to himself in conventionally polite, self-effacing phrases like "in the most lowliest wyse that I can, I mekely recomend me unto you", portraying himself as a respectful son-in-law. On the other hand, John Holles in a letter to his father (2) comes across as self-confident, even arrogant, complaining in the first person about his small allowance. Self-reference in the first person is, of course, only one available way for the writer to position him/herself; even the self can be referred to by other means, like third-person singular or first-person plural pronouns, noun phrases (e.g. *your good son*, *your most obedient son*, *your friend*, *your humble servant*) or proper names.

Self-reference follows genre-specific conventions. Dialogic genres like face-to-face conversations or personal letters contain frequent self-reference as they are communication between the writer (*I*) and the recipient (*you*). Thus, first-person singular pronouns can be regarded as prototypical self-reference in personal correspondence. In other genres, like present-day academic writing intended for a wide audience, self-reference is more unusual, but it is still an important resource for the writer to construct an appropriate authorial self as a scholar (e.g. Hyland 2001; Samson 2004). However, the frequency of self-reference varies from one situation to another, even in personal correspondence, depending on the writer-addressee relationship. For instance, letter writers tend to refer to themselves more often when writing to close and socially equal recipients than when writing letters to more distant or socially superior addressees (e.g. Palander-Collin 2006, 2009a; Nurmi and Palander-Collin 2008).

In this paper, I further explore the role of self-reference (*I*) in sixteenth- and eighteenth-century family and non-family letters. Family letters should represent a more intimate and relaxed style, while letters to acquaintances should show a more public style. In order to see how the writers position themselves, quantitative corpus methods are used; and to make some sense of the significance of the varying and increasing frequencies of self-reference observed in sections 4 and 5, a functional approach to self-referential mental processes is adopted in section 6. The analysis reveals three tendencies: i) the frequency of *I* increases from the sixteenth to the eighteenth century, ii) mental processes emerge as the central process type in the first person, and iii) self-referential mental processes often serve interpersonal functions. This analysis also confirms the earlier finding that self-reference is more frequent in family correspondence than in letters to more distant acquaintances. These findings are placed in the context of results obtained from earlier studies concerning increased stance marking, involvement and subjectification in various other genres in the history of English from around the seventeenth century onwards (Biber and Finegan 1989; Biber 2004; Taavitsainen 2002). The deeper causes of such changes – that is if the tendencies are related – remain to be explored more fully, but it seems that they may reflect broader ideological changes in ways of thinking, argumentation and building rapport with the reader.

2. Pronouns as indexicals

The idea of the first person as a linguistic feature relating to identity and relational functions of language can be based on Mühlhäusler and Harré's (1990: 87-105) thesis of double indexicality for *I*, by which "I and other first-person expressions are used not to denote anything, but as indices of location" (Mühlhäusler and Harré 1990: 92). These locations are of two kinds as *I* indexes the speaker's utterance i) in place and time and ii) with the person to be held morally responsible for it. Mühlhäusler and Harré (1990: 94-95) suggest that speakers are always located in both ways, physically and morally, and that physical location tends to be universal, stable and independent of cultural conventions, whereas moral geography is local, variable and embedded in different registers. The location of speakers in the moral order of speaking is defined by various permanent as well as temporary factors. More or less permanent factors include the rights, duties and obligations of the speakers and hearers and relate to their social roles in the context, while temporary factors refer to the moment-by-moment presentation of self as confident, shy, competent, arrogant, and so on.

This model, like, for example, Layder's (2003) social theory, incorporates the macro- and microlevels of analysis within the same framework in recognizing on the one hand that there are macrosocietal conventions and categories, and on the other hand that writers make meaning and strategic choices in local contexts, like being arrogant towards their father. Layder (2003: 2-5) suggests that there are four domains in social life – psychobiography, situated activity, social settings

and contextual resources – that relate to each other and are always present. These domains situate individuals in a complex web encompassing personal feelings, attitudes and predispositions, the current face-to-face transaction, the particular setting of a specific location or social organization, and the society-wide distribution and ownership of resources along gender, race and class lines.

Socio-historical research helps us create an understanding of the more permanent factors defining the moral order of the period and its people. At a minimum, a person's social position and gender are important identity and relational factors, as England was a highly stratified society where people of different rank and gender participated in different activities and had different legal rights. Wrightson (2003: 25), for example, writes that in early modern England "the most fundamental structural characteristic of English society was its high degree of stratification, its distinctive and all-pervasive system of social inequality". The separation of civil and domestic life was also evident (Bryson 1998: 167), as shown in the style-shifting of examples (3) to (6). Different types of language were advised in letter-writing manuals and conduct books of the period to address people of various ranks (Bryson 1998: 166-168). According to manuals, superiors should, for instance, be allowed the initiative in discourse, while inferiors should avoid blunt questions and direct contradiction. In response to a superior, the inferior should use the appropriate title, avoid direct address with the pronoun *you*, insist on the superior's freedom from any obligation to the inferior, and show submission. These features are evident in examples (4) and (6), which employ circumlocution and repeat the title *Your Lordship/Grace* at intervals. Although the nature of appropriate self-reference as such was not necessarily regulated, the normative ideas express the expectations placed on a gentleman's verbal behaviour and clearly show that a varying portrayal of the self was the prescriptive norm, depending on the situation.

(3) Nann your sick sister hath made one and a most earnest request, that she might see you, which she saith would exceedingly comfort her: as *I am loath* you should cum hither, being a sick house, so *am I willing* to satisfy her, and the rather because *I fear me* you shall not see her recovered, her sickness grows so fast upon her: sweet Nan for those reasons *I would have you cum*, but *long* you shall not stay, unless you might have more comfort in staying: for this time wishing you no less good then myself, farewell, *I am your most loving husband John Holles*. (CEEC: John Holles to his wife, Ann Holles, 1591: p. I, 2)

(4) My Lord *I fynd* your Lordship according to your place, allwayes so accumpanied with business, as without injurie to your Lordship or prejudice to my self, *I can not have*, or *take time* for more words, then serves for salutation, which causeth me to resort to this paper, uppon which your Lordship may cast your ey at your pleasure, least my desires, in attendance of other opportunity, still smothered in my owne brest, may have juster cause to blame my self, for their is not satisfaction, then either frends or fortune […] in the mean time *I humbly take my leave and rest*. Your Lord-

ships ever ready to serve yow J. Holles. (CEEC: John Holles to the Earl of Somerset, 1615: p. II, 61)

(5) My dear Sam
I write to you in great haste rather to apologise for not writing, if *I may say* so without an Iricism, than having time to write to any other purpose. the course of conveyance between us is so precarious, that though *I answer'd* your's as soon as 'twas possible for me and you replied to mine as soon as 'twas possible for you, your last letter did not get to town before yesterday, nor to my hands before this morning. and now it has got to my hands it is at a time when unfortunately *I cannot pay* that attention to it *I cou'd wish and would have done had I not been* under an engagement for the remainder of the day of long standing. while *I am now writing* there is actually a Gentleman in the room upon a first visit to me, of whom *I have begged* permission to retire to my desk for a few minutes to give you as many lines. (CEECE: Jeremy Bentham to his brother Samuel Bentham, 1769: p. I, 136)

(6) My Lord
The liberty *I took* of sending to Lord Spencer the letter with which your Grace was pleased to honour me would *I hoped* have spared me the necessity of being any further troublesome to your Grace on that score. By a passage in a subsequent letter of his Lordship which *I inclose* it appears that in that expectation *I was* rather too sanguine. *I find* myself obliged therefore to sollicit at your Grace's hands a few words of explanation addressed to his Lordship to satisfy that 'concurrence and consent on your Grace's part of which he expresses himself not yet sufficiently assured, and without which he declares it impossible for him to proceed a single step.' (CEECE: Jeremy Bentham to Archbishop William Markham, 1793: p. IV, 471)

Example (3) comes from John Holles's (1565?-1637) letter to his wife, and example (4) from a letter to his patron Robert Carr, Earl of Somerset. The role of husband and someone seeking patronage from a social superior are reflected in the language as conventional choices (e.g. "Nann" versus "My Lord" as appropriate forms of address), but also in self-reference. Writing to his newly-wed wife about her sister's possibly fatal illness, the "loving" husband repeatedly expresses his own attitudes, wishes, fears and instructions to the wife in the first person, whereas when seeking patronage from the Earl of Somerset, Holles takes a considerable amount of space to approach Somerset modestly and indirectly before eventually voicing his worries. Self-reference (*I*) is also more sparingly used and serves to compliment the Earl. These examples show self-reference both as an outcome of larger societal processes and structures, and as local interactional negotiation that is partly conventional and habitual and partly less than fully conscious (see Bucholtz and Hall 2005 on identity construction).

Examples (5) and (6) come from comparable contexts roughly two centuries later and show similar style-shifting between familial and civil styles. They

were also written by a gentleman, philosopher Jeremy Bentham (1748-1832), to his brother (5) and to his father's friend and his own old headmaster, now Archbishop of York, about a controversial purchase of land owned by the see of York (6). Self-reference is used several times and Bentham refers to his feelings and attitudes in both examples. The first-person expressions show his disposition towards the recipient: he is a caring elder brother in (5) and an appropriately respectful gentleman in (6).

This study focuses on gentlemen's linguistic choices, as it can be assumed that men from other social spheres, or women, would have different access to various styles. Moreover, gentlemen were literate, and plenty of letters written by them have been preserved and edited. However, the category of gentleman is somewhat difficult to define as the boundaries of the gentry were not legally stated (e.g. Heal and Holmes 1994; Porter 1990; Rosenheim 1998; Wrightson 2003). Lineage, wealth and landed property were important material characteristics of a gentleman, but also moral virtues such as intellectual interests, Christian beneficence for the poor and engagement with rural pastimes were associated with the gentry (Heal and Holmes 1994: 277). Gentlemen were usually well-educated and they attended the universities and the Inns of Court and were regularly sent on the Grand Tour. Many gentlemen held influential public posts and belonged to the ruling class. In this study, titled aristocracy was excluded although many of the informants were upwardly mobile and were eventually raised to the peerage, like John Holles cited above (Seddon 2006). Heal and Holmes (1994: x) also argue that the status of the gentry changed in important ways from 1500 to 1700, as it was increasingly integrated into a national culture through education, increased mobility and urbanization. Such changes probably affected the moral landscape as well as the language. It is, nevertheless, difficult to prove connections between societal changes and linguistic changes, but I believe that links exist.

3. Data and methodology

The data for this study comes from the *Corpus of Early English Correspondence* (CEEC) and the *Corpus of Early English Correspondence Extension* (CEECE). The sampling focused on sixteenth- and eighteenth-century letters written by men who can broadly be classified as members of the gentry. Letters addressed to family members and non-family acquaintances were selected (table 1).[1]

Two types of quantitative corpus-based analyses were carried out to establish patterns of self-reference in the data. First, the occurrences of *I* in the small sample were analyzed for the semantic type of the main verb (as well as modal and auxiliary elements in the verb phrase). Second, WordSmith (Scott 2004-2007) was used to extract two-word clusters containing *I* in the big sample in order to see which individual lexical verbs emerge as the most frequent ones in self-reference. Clusters with a frequency of at least 0.1 in 1,000 words were considered if they were used by at least three writers. Such a cluster analysis

produces a picture of recurrent patterns, but does not describe the data exhaustively. Statistical significance was calculated with chi-square tests, and the five percent level was considered significant throughout.

Table 1: Samples from the *Corpus of Early English Correspondence* (CEEC) and its *Extension* (CEECE)

	CEEC 16th century		CEECE 18th century	
Word count	Non-family	Family	Non-family	Family
Big sample	127100	120800	62500	74100
Small sample	19394	12847	10085	7235
Informants	13	13	9	6

For data analysis, the two samples of different sizes were created since even a small sample yields a considerable number of *I*'s. The general frequencies of the pronoun *I* and the two-word *I*-clusters were calculated from the big sample, but the more time-consuming verb type analysis was based on the small sample.[2] Letters for the small sample were randomly selected from all the writers included in the big sample; two or three letters were included from each writer. I also employed the big sample for another cluster analysis in Palander-Collin (2009b); for this purpose, the spelling of the material was standardized and the standardized material was employed here. However, the examples are given in their original (i.e. editorial) spelling. In comparison to the results obtained from the big sample, the validity of the small sample is statistically adequate (see table 2 below).

4. Frequency of self-reference

The use of the first-person subject *I* shows synchronic variation and diachronic change in terms of frequency. Table 2 indicates that self-reference in gentlemen's letters increased from the sixteenth to the eighteenth century, and that the family context favoured self-reference more than correspondence between non-family acquaintances. The register and diachronic differences shown in table 2 are statistically significant.

The increase in self-reference in gentlemen's letters may relate to increasing involvement and stance marking observed in the history of English. Biber and Finegan (1989) showed that at least fiction, essays and letters in the ARCHER corpus have undergone a diachronic drift towards more involved styles from 1650 up to the present. First-person pronouns are an involvement feature like, for instance, second-person pronouns, the present tense, private verbs and contractions. Moreover, Biber (2004: 107) claims a steady increase in stance expressions showing speakers' or writers' "epistemic or attitudinal comments on propositional information" in drama, letters, newspapers and medical prose during the same period. Stance expressions include a wide variety of linguistic forms from

modals and stance adverbials to various complement clauses and complement taking predicates such as *I think* (Biber 2004: 112; Englebretson 2007: 17). From the point of view of language change, popular registers of drama and personal letters lead the way in the increased use of stance markers (Biber 2004: 130). The results of the current analysis also suggest that more informal and intimate styles lead the increase in the frequency of self-reference in gentlemen's letters. Increasing frequencies of self-reference have also been observed in seventeenth-century medical writing, where argumentation in the first person as well as the use of mental verbs seem to correspond to the rise of the empirical paradigm and new empirical ways of producing and presenting knowledge instead of the old scholastic thought-style relying on the writings of ancient authorities as evidence (Taavitsainen 2002).

Table 2: The absolute and normalized frequency (per 1,000 words) of *I* in gentlemen's non-family and family letters in the sixteenth and eighteenth century

	16th century		18th century	
	Non-family	Family	Non-family	Family
Big sample	2852	3478	1968	2555
	22.4	28.8	31.5	34.5
Small sample	424	360	291	268
	21.9	28.0	28.9	37.0

5. Self-referential verbal processes

I have approached the nature of self-reference in terms of semantic verb types attached to the first-person subject as they are indicative of the process types favoured in various contexts. Present-day English recipes, for instance, tend to employ material processes, narratives express existential and relational processes, while casual conversations exhibit mental processes (Halliday and Matthiessen 2004: 174). Moreover, Halliday and Matthiessen (2004: 170) claim that one of the most basic differences that we become aware of at a very early age is the one between an inner and outer experience: "between what we experience as going on 'out there', in the world around us, and what we experience as going on inside ourselves, in the world of consciousness (including perception, emotion and imagination)". The outer experience concerns actions and events; things happen, people or other actors do things or make them happen. The inner experience reflects and reacts to the outer experience. Prototypical outer experiences are grammatically expressed as material (transitive) processes (e.g. *people make money*), while prototypical inner experiences are mental processes (e.g. *people love money*).[3]

5.1 Semantic verb types and the verb phrase

This analysis employs the semantic verb types presented in Biber et al. (1999: 360-363), who divide verbs according to their meaning into activity, communication, mental, causative, occurrence, existence/relationship and aspectual verbs. The activity, communication, existence/relationship and mental verbs are relevant for this discussion, as the other types are only marginally attested in the data (table 3). Activity verbs refer to volitional activity where an action is performed intentionally by an agent or 'doer' (Biber et al. 1999: 361), and they correspond to a great extent to Halliday and Matthiessen's (2004) material processes expressing outer experiences. Communication verbs are a subcategory of activity verbs that involve communication activities (Biber et al. 1999: 362). Existence/relationship verbs in the data are mainly the main verbs BE and HAVE, and mental verbs refer to mental states and activities experienced by humans (Biber et al. 1999: 362-363). Quirk et al. (1985: 1180-1182) call these verbs "private" as they are not open to observation.

Table 3: Frequency of semantic verb types with the subject *I* (relative frequency, absolute frequency and frequency per 1,000 words)

	16th century		18th century	
Verb type	Non-family	Family	Non-family	Family
Activity	21% 91 / 4.7	17% 62 / 4.8	15% 45 / 4.5	18% 49 / 6.8
Communication	24% 100 / *5.2	30% 108 / *8.4	19% 55 / 5.5	20% 54 / 7.5
Existence/ relationship	17% 71 / 3.7	18% 63 / 4.9	24% 70 / *6.9	26% 71 / *9.8
Mental	33% 142 / 7.3	33% 117 / 9.1	40% 117 / 11.6	34% 90 / 12.4
Other	5% 13 / 0.7	3% 10 / 0.8	1% 4 / 0.4	2% 4 / 0.6
Total	100% 424 / *21.9	100% 360 / *28.0	100% 291 / *28.9	100% 268 / *37.0

* Significant non-family versus family difference at five percent level

The analysis, thus, focuses on lexical verbs in the verb phrase. Simple verb phrases are the prominent type in the data, covering approximately 60 percent of the first-person verb phrases. Approximately 22 percent of the verb phrases contain a modal auxiliary and 18 percent are formed with the primary auxiliaries BE, HAVE or DO.[4] The overall frequencies are surprisingly stable and no statistical

differences emerge diachronically or between the registers. Different verb types, however, behave quite differently (figure 1). Activity verbs occur mostly in complex verb phrases with e.g. aspectual or modal marking, while mental and existence/relationship verbs favour the simple present and the past tense. Such relationships seem to be fairly stable, as the semantic quality of the verb types either attracts or does not attract aspectual or modality marking or the passive voice. It is also possible to observe a change in the verb phrase structure favoured by communication verbs, but rather than being a typological shift, the decrease in simple verb phrases seems to reflect a decrease in request and politeness markers formed with communication verbs (see 5.2 below). Hence, a structural analysis makes sense in functional and interpersonal terms.

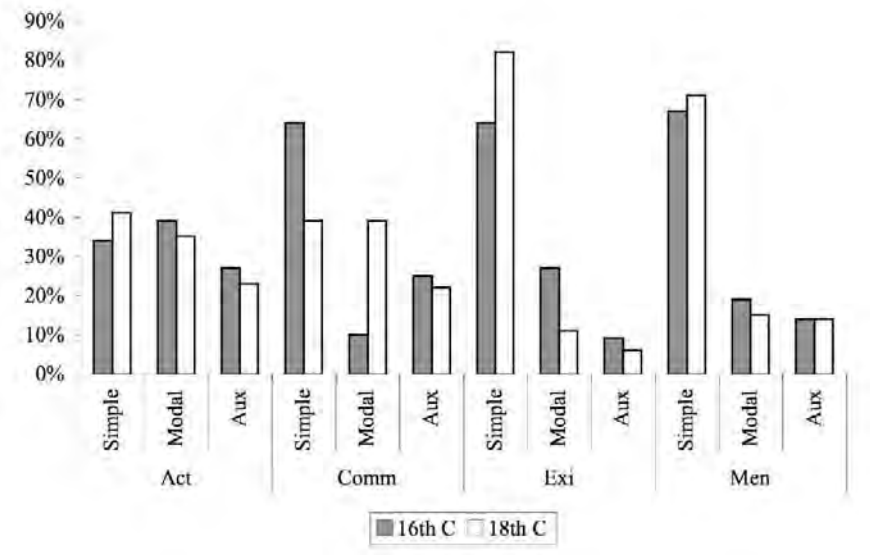

Figure 1: The structure of the verb phrase with activity, communication, existence/relationship and mental verbs in the sixteenth and eighteenth century

5.2 Frequency of semantic verb types

Table 3 above shows the frequencies of semantic verb types used with the first-person subject *I*. Family letters consistently contain more occurrences of self-reference with each verb type, but these register differences are statistically significant only in the case of communication and existence/relationship verbs (marked with * in table 3). Thus, it seems that register differences are difficult to describe further with this apparatus and a more contextual and detailed reading is needed. Figure 2, however, illustrates a diachronic change, showing that mental

verbs and existence/relationship verbs are significantly more frequent in the eighteenth century than in the sixteenth century. The increase in self-reference observed in section 4, then, occurs most clearly in these process types. In all respects, mental verbs emerge as the largest category in personal letters, covering over 30 percent of all processes attached to the first-person subject. Mental verbs will be discussed in more detail in sections 5.3 and 6, while other verb types are briefly described here.

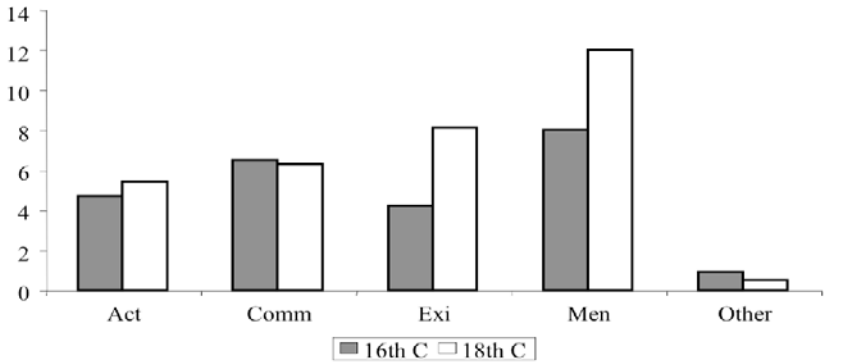

Figure 2: Frequency (per 1,000 words) of semantic verb types with the subject *I* in the sixteenth and eighteenth century

Communication verbs emerge as the second biggest category after mental verbs in the sixteenth century, whereas existence/relationship verbs are the second largest category in the eighteenth century. It seems that in sixteenth-century letters, communication verbs are an important source of opening and closing formulae like *I (re)commend me unto you*, *I commit you to God* and *I bid you farewell* (examples (7)-(9)); also many other frequent first-person formulae were formed with communication verbs like the request markers *I pray* and *I beseech*, and the attitude marker *I thank God* (examples (10)-(12)). These no longer emerge prominently in the eighteenth century, which was also apparent in the decreasing frequency of simple communicative verb phrases in figure 1. Apart from these formulaic expressions, communication activities in the first person are often about writing, sending or receiving letters, but also about saying (example (12)) and thanking.

(7) Master Derbye, *I hertely commend me unto you*, and wheras you have sene and herd divers per*sons* afor you redy to averre Robert Waltons sklanders *by* hym spoken and utered to my wyfis dishonestye *and* myn wyche sklanders they wer redy to justefye afor Waltons face. (CEEC: Sir Edward Willoughby to Thomas Darby, 1530?: p. 82)

(8) All your frends at Fulham are well. I was there on Monday last, and so wishing to you and yours *I commit you to God*. From London this 5th of Aprill 1606. Yours most assuredly John Chamberlain. (CEEC: John Chamberlain to Ralph Winwood, 1606: p. I, 226)

(9) Thus with my commendations to your wyf & other my brothers and sisters, *I bid you farewell*. (CEEC: Edward Bacon to his brother Nathaniel Bacon, 1577: p. I, 280)

(10) And *I pray* you let my wyfe have some word from you by this next carryer, how you doe in your sayd matters; and she and my sister Ellynor humbly recomend them unto you, and pray you for your dayly blessing. (CEEC: Germain Pole to his father-in-law Sir Robert Plumpton, 1502: p. 166)

(11) *I beseech* you pardun mi rudenes if I do otherwise then I shuld do; and blame their unkindnes that use me otherwise then they can or ouht to do. (CEEC: Gabriel Harvey to John Young, Master of Pembroke Hall, Cambridge, 1573: p. 1)

(12) Your sister knew not my sending, else yow had a long epistle from her, for wemen to wemem never want discours: shee is *I thank God* reasonably well; yet her head-ake still continuing, *I may say*, though I hope better, shee doth trainer son lien: God grant her a good recovery, that I may see her in the world, as yow ar, (CEEC: John Holles to his daughter Arabella Wentworth, 1631: p. III, 431)

The increase in the frequency of existence/relationship verbs particularly concerns the verb BE and relates partly to the frequent use of the closing formula *I am*... in the eighteenth century letters (example (13)). However, stance phrases like *I am glad* (14) and *I am sorry* and (15), are also increasingly used in the eighteenth century. They are semantically comparable to mental verbs in that they show the writer's attitude or state of mind, and Halliday and Matthiessen (2004: 212), for example, consider that these relational processes, like mental processes, construe inner experiences as opposed to outer experiences. Various *I am*... phrases are also linked to the increase in the simple existential verb phrase shown in figure 1 above.

(13) *I am*, Sr Your most humble Servant John Gilbert-Cooper. (CEECE: John Gilbert Cooper to Robert Dodsley, 1745: p. 90)

(14) *I am very glad* the good Bp of Gloucester will be so soon in town I hope He will arrive in good Health. (CEECE: Brown Willis to Thomas Secker, 1741: p. 70)

(15) *I* have been misinformed, and *am sorry* for the mistake. (CEECE: William Jones to the first Earl Spencer, 1770: p. I, 75)

Letter writers describe their activities less frequently than could perhaps be expected, and activity verbs do not emerge as prominently as other verb types. The most frequent activity verbs in the first person relate to coming and going,

taking and leaving, getting, meeting and making. Some activity verbs also occur in idiomatic phrases like *take the liberty* in the eighteenth century, or in the closing formulae *I take my leave* or *I leave you to God* in the sixteenth century.

5.3 Mental verbs

Halliday and Matthiessen (2004: 210) divide mental processes further into perceptive (e.g. *perceive*, *sense*, *see*, *hear*), cognitive (e.g. *think*, *believe*, *doubt*, *remember*), desiderative (e.g. *want*, *wish*, *desire*, *hope*) and emotive (e.g. *like*, *love*, *fear*, *regret*, *marvel*) processes. Using this categorization of mental verbs, we notice that the cognitive type, with common evidential/epistemic expressions *I think* and *I know*, dominates in both time periods and registers, covering approximately 50 percent of self-referential mental verbs. The frequency of all the subcategories of mental verbs, except perceptive verbs, increases over time (figure 3).

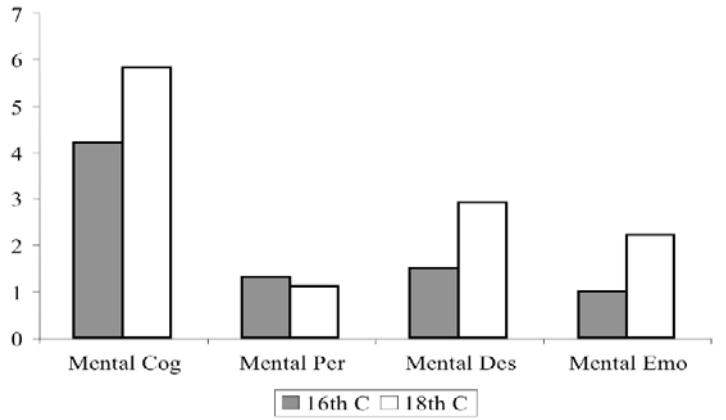

Figure 3: Frequency (per 1,000 words) of subtypes of mental verbs with the subject *I* in the sixteenth and eighteenth century

As to the increase in the use of desiderative verbs, it seems that politeness marking shifts from the use of *I* + communication verb (e.g. *I pray*, *I beseech*, *I commit*, *I bid*, *I assure*) to the pattern *I* + desiderative mental verb. Desiderative verbs in the eighteenth century are used to express the writer's concern and wishes for the recipient's wellbeing (example (16)), but also as request markers (example (17)).

(16) and *I hope* you'l continue to recover sensibly every day, which *I heartily wish*. (CEECE: Thomas Yorke to his brother-in-law Sir James Clavering, 1724: p. 150)

(17) *I desire* you would still remind the Printers not to reject the other mark. (CEECE: John Gilbert Cooper to Mr Dodsley, the bookseller, 1746: p. 105)

Table 4: The most frequent *I* + mental verb clusters (freq. ≥ 0.1 per 1,000 words). Type: P=perceptive, C=cognitive, D=desiderative, E=emotive

		16th century				18th century			
		Non-family		Family		Non-family		Family	
Cluster	Type	No.	Fq	No.	Fq	No.	Fq	No.	Fq
I hope	D	*33	0.3	*79	0.7	*62	1.0	*104	1.4
I think	C	*40	0.3	*73	0.6	44	0.7	70	0.9
I believe	C					*21	0.3	*51	0.7
I wish	D			24	0.2	32	0.5	48	0.6
I know	C	*41	0.3	*68	0.6	24	0.4	28	0.4
I find	P	22	0.2	26	0.2	23	0.4	25	0.3
I suppose	C	29	0.2			9	0.1	24	0.3
I fear	E			13	0.1			21	0.3
I found	P					7	0.1	14	0.2
I hear	P	18	0.1	23	0.2			14	0.2
I mean	C			35	0.3	8	0.1	14	0.2
I desire	D					9	0.1	11	0.1
I saw	P					10	0.2	10	0.1
I thought	C	29	0.2	22	0.2	9	0.1	10	0.1
I dare	E					13	0.2	8	0.1
I understand	C							8	0.1
I doubt	C	*24	0.2	*41	0.3	10	0.2		
I trust	E	*18	0.1	*34	0.3	10	0.2		
I heard	P			20	0.2				
I see	P	24	0.2	18	0.1	11	0.2		
I knew	C					9	0.1		
I perceive	P	14	0.1						

* Significant non-family versus family difference at five percent level

Table 4 above summarizes the output of WordSmith cluster analysis for the most frequent two-word *I* + mental verb clusters. Many of the clusters occur in both time periods and registers, but interesting differences can also be observed. For instance, the sixteenth-century letters produced fewer clusters than the eighteenth-century letters, although the sample size was considerably bigger for the sixteenth century. *I* + mental verb clusters tend to be more common in family correspondence just like *I* + mental verb combinations in general. Many individual clusters are significantly more frequent in family letters, particularly in the sixteenth century, including the top three clusters *I hope*, *I think* and *I know* (significant differences marked with * in table 4). Thus, it seems that these expressions stem

from more informal contexts characterized by authorial presence and the involvement between the writer and the recipient.

Focusing only on the topmost *I* + mental verb clusters *I hope*, *I think*, *I believe*, *I wish*, *I know* and *I find*, figure 4 shows graphically the increase in their frequencies from the sixteenth century to the eighteenth century. The increase is statistically significant except for the cluster *I know*.

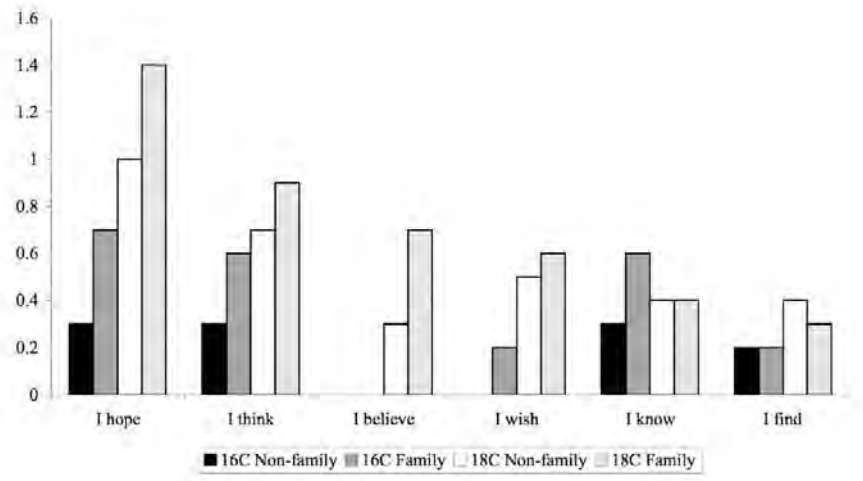

Figure 4: Frequency (per 1,000 words) of the topmost *I* + mental verb clusters in sixteenth- and eighteenth-century non-family and family letters

6. Functions of mental verbs

It seems to me that the increase in the use of the first-person subject and *I* + mental verb patterns cannot be explained without reference to their communicative functions. For one thing, the increasing use of the pronoun *I* with mental verbs seems to indicate that the letter-writing styles became more subjective over time, but such subjectivity actually often turns out to be intersubjective in nature, highlighting the interpersonal (politeness) work between the *I* and *you* of correspondence. In this section, I discuss the role of self-referential mental verbs as interpersonal devices in the stance framework, as I believe that changes in patterns of self-reference can be understood as shifts in ways of expressing attitudes and alignments, although further work in this area is definitely needed.

In broad terms, we can say that self-referential mental verbs potentially serve as stance markers in that they express "personal feelings, attitudes, value judgements, or assessments" (Biber et al. 1999: 966). Englebretson (2007: 6) presents five principles of stancetaking that relate stance to interaction between the participants so that stance is collaboratively constructed with respect to others

and it is indexical, evoking aspects of broader societal structures. Du Bois (2007) discusses three aspects of a stance act, which are seen as subsidiary rather than separate. These are evaluation, positioning and alignment. In stancetaking these aspects have their own distinctive consequences, as Du Bois (2007: 163) writes: "In taking a stance, the stancetaker (1) evaluates an object, (2) positions a subject (usually the self), and (3) aligns with other subjects". Not all the aspects of stance are always specified, but participants may still be able to make inferences about the ones not mentioned (Du Bois 2007: 164). For example, by uttering *I agree* the speaker normally aligns with the hearer, taking the same stance with him/her and including the same evaluation as the hearer has previously made.

Here, I shall briefly discuss the functions of *I* + mental verb sequences in letters, focusing particularly on the most frequent two-word clusters *I hope*, *I think*, *I believe*, *I wish*, *I know* and *I find*. Employing the three aspects given by Du Bois, we can say that the two-word clusters are complement-taking predicates evaluating "objects" in the complement clause. By writing the sequence *I think*, *I believe*, *I know* or *I find* the writers position themselves on an epistemic scale evaluating the certainty or uncertainty of the proposition and their commitment to it, while *I hope* and *I wish* perhaps position the writers more along an affective scale, expressing desires and wants. The alignment in the personal letters often seems to be with the recipient, but in the following I shall explore various possibilities that highlight the interactional nature of *I* + mental verb clusters.

In examples (3) to (6) above we saw that self-reference among other linguistic features creates familial and civil styles that are used according to the writer-recipient relationship. The alignment of self-reference is another salient characteristic in its use. In some cases, self-reference seems to focus on the writer, his feelings and attitudes, but in other cases it is clearly used to relate the writer to the recipient, either to attend to the recipient's needs, such as with a compliment, or to make sure that he or she will be favourably disposed towards the writer. Alignment may change within a letter, but it may be that the writer's needs in general are given more space in family letters, while the recipient's needs have to be more attended to in non-family letters. Hierarchical status differences are another possible factor affecting alignments, and it may be assumed that the writer's needs receive more space between equal correspondents. These ideas follow Dressler and Wodak's (1982) proposition that in shifting from informal to formal situations the reader's/hearer's needs override those of the writer/speaker.

Examples (18)-(23) illustrate various alignments in sixteenth-century data with *I hope*. In example (18), the writer anticipates a happy meeting with his brother, which can be regarded as a compliment towards the recipient, making him feel good. In the same way, writers often hope to see the recipient or hear from him/her soon or again. In (19) and (20) the writer's hopes are again directed towards the recipient, but this time in order to affect the recipient's sympathy towards the writer and his conduct. In (21) the writer expresses a hope concerning his brother's conduct, as the elder brother is in the position of selling an estate to cover his debts, something the younger brother would not want him to do.

(18) And now let me conclude with the best and most faythfull affection that ever any brother colde beare to his brother, And *I hope* that God will send vs a happy meeting about some fower months hence. (CEEC: Philip Gawdy to his brother Bassingbourne Gawdy Junior, 1591: p. 61)

(19) I recevved the mony yow sent vpp, the Somm of 5^{li}. Suerly, Sir, I have had great occasion to vse mony, w^{ch} hathe made me vse my creditt. And *I hope* you will not dyslike of my dealing therin, for I have sufficyent to show and the parties be very sufficyent. (CEEC: Philip Gawdy to his father Bassingbourne Gawdy Senior, 1587: p. 29)

(20) My promise to your lordship touching the sending of a general note of your mineral debts I hasten to perfourme; and now shal your lordship receyve the same, and if it be not perfite *I hope* you wil beare with me and upon notice geven I wil labour to refourme it. (CEEC: Francis Hastings to his elder brother Earl of Huntingdon, 1585: p. 32)

(21) Sutes be lingering and uncerteyne, and your lande is so farre spente allready as (under your lordship's correction be it spoken) I know not where it wilbe had; for if only Ringwod goe for Atterton's debt (which I doubt it wil not reache) yet is Mr. Hare to be satisfyed a rounde sume from your lordship for yourselfe, and lande must goe for that and your other debts, and only Christchurche and Canforde have you in these partes to sel – for *I hope* you wil never so muche as once imagine of the sale of Aller. (CEEC: Francis Hastings to his elder brother Earl of Huntingdon, 1592: p. 50)

Example (22), on the other hand, seems to be genuinely oriented towards the writer, who for his own sake hopes that he will not regret his actions. In example (23) the writer's hope is directed to a third party and it is not in any apparent way related to the recipient. The hope, moreover, seems to be conventional rather than sincere.

(22) I have not as yeat repented the first motion of this jorney, & *I hope* God will give me a mynde verie constant therin. (CEEC: Edward Bacon to his brother Nathaniel Bacon, 1576: p. I, 203)

(23) I must pray your Lordship to excuse me that I have ben silent so long, having receved so many kind letters from you, but indeed I have not, nor was not able to set pen to paper many a day, but only about a fortnight since for a farewell to Master Gent, who goode man is gon to God *I hope*. (CEEC: John Chamberlain to Sir Ralph Winwood, 1613: p. I, 446)

As the interpersonal function and alignment towards the recipient seem to be central characteristics of *I hope* as well as other *I* + mental verb clusters, I checked whether this interpretation could be further tested using corpus methods. For this purpose, I counted how often the recipient, usually in the form of the pronoun *you* or a title like *Your Lordship*, is actually mentioned in the same sentence with the cluster so that the self-referential mental process consequently

relates to the recipient in one way or another. Although this analysis does not produce a detailed understanding of how the recipient relates to the *I* + mental verb cluster, it shows that the recipient is indeed frequently mentioned: in non-family letters the addressee is mentioned in the connection of *I* + mental verb cluster in some 40 percent of cases, while in family letters the addressee is mentioned in some 52 percent of cases.

7. Conclusion

This paper set out to explore changing patterns of interaction using corpus data and methods. The analysis focused on self-reference in sixteenth- and eighteenth-century letters written by gentlemen. Self-reference was understood as a linguistic feature relevant to the identity and interpersonal functions of language. The results indicate that in gentlemen's letters self-reference increased from the sixteenth to the eighteenth century. As family letters contained more frequent self-reference than non-family letters, it seems that the linguistic practices of the more informal and intimate context spread to other, more formal, contexts.

The increase in self-reference may relate to changes in broader patterns of stance marking, involvement and subjectification as they have been found to increase at least from 1650 onwards (Biber and Finegan 1989; Biber 2004; Taavitsainen 2002). Mental verbs were identified as the most frequent semantic verb type governed by the first-person subject, and the frequency of mental verbs together with existence/relationship verbs increased from the sixteenth to the eighteenth century. It seems that the increase in the frequency of self-reference may be explained in terms of an increasing emphasis on the writer's internal mental processes, as even existence/relationship verbs are regularly used as expressions of feelings and attitudes (e.g. *I am glad*, *I am sorry*). The increase in self-referential mental processes further relates to patterns of interaction: *I* + mental verb sequences often serve the writer's and recipient's needs in various ways and build rapport, in particular with the reader.

Further work could be done on self-reference. The connection between self-referential mental processes and stance marking deserves further research. First, the interpersonal alignment of *I* + mental verb sequences could be analyzed in more detail to see whether any changes take place in the ways in which the writer positions himself in relation to the recipient or other persons. Second, the analysis of two-word clusters produced quite a few (semi-)fixed *I* + mental verb combinations with increasing frequencies. It may be hypothesized that with increasing stance marking in general, a number of first-person expressions were routinized or even grammaticalized to signal stance in dialogic genres. Finally, deeper causes of stylistic shifts and changes in patterns of interaction described here and their links to broader ideological or societal shifts remain to be explored.

Notes

1 The sixteenth-century informants include: Germayn Pole (F), Thomas More (NF, F), Thomas Boleyn (NF), Edward Willoughby (NF), Thomas Elyot (NF), Thomas Clifford (F), Thomas Wyatt (NF, F), Thomas Wharton (NF), Nicholas Bacon II (F), Nicholas Bacon I (NF, F), Edward Bacon (F), Gabriel Harvey (NF, F), Francis Hastings (NF, F), Edward Clere (NF), Thomas Gresham (F), John Holles (NF, F), Philip Gaudy (F), Thomas Edmondes (NF), Robert Cecil (NF); The eighteenth-century informants include: Thomas Yorke (F), John Yorke (F), William Pulteney (NF), Browne Willis (NF), Andrew Stone (NF), Edward Gibbon (F), Jeremy Bentham (NF, F), John Gilbert Cooper (NF), Pierce Taylor (F), Samuel Crisp (F), William Jones (NF), Reginald Pole Carew (NF), George Canning (NF). NF = non-family letters, F = family letters.

2 I am grateful to Mia Kylmälä for helping me with the verb type analysis and the Sociocultural Reality and Language Practices (SoReaL) project funding granted by the University of Helsinki that made it possible for her to be employed.

3 Other process types include behavioural processes (e.g. *people are laughing*), verbal processes (e.g. *so we say*), relational processes (e.g. *time is money*) and existential processes (e.g. *there's Christianity in the South*) (see Halliday and Matthiessen 2004: chapter 5 for more details of this model).

4 Some complex verb phrases contained both modal and primary auxiliaries, but there were not many and they were, therefore, simply categorized as modal.

Corpora

Corpus of Early English Correspondence (CEEC). Compiled by T. Nevalainen (leader), J. Keränen, M. Nevala (née Aunio), A. Nurmi, M. Palander-Collin and H. Raumolin-Brunberg, http://www.helsinki.fi/varieng/domains /CEEC.html.

Corpus of Early English Correspondence Extension (CEECE). Compiled by T. Nevalainen (leader), S. Kaislaniemi, M. Laitinen, M. Nevala, A. Nurmi, M. Palander-Collin, H. Raumolin-Brunberg, A. Sairio (née Vuorinen) and T. Säily, http://www.helsinki.fi/varieng/domains /CEEC.html.

References

Auer, P. (2007), 'Introduction', in: P. Auer (ed.) *Style and Social Identity. Alternative Approaches to Linguistic Heterogeneity*. Berlin and New York: Mouton de Gruyter. 1-21.

Biber, D. (2004), 'Historical patterns for the grammatical marking of stance. A cross-register comparison', *Journal of Historical Pragmatics*, 5(1): 107-136.

Biber, D. and E. Finegan (1989), 'Drift and the evolution of English style: a history of three genres', *Language*, 65(3): 487-517.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bryson, A. (1998), *From Courtesy to Civility: Changing Codes of Conduct in Early Modern England*. Oxford: Clarendon Press.

Bucholtz, M. and K. Hall (2005), 'Identity in interaction: a sociocultural linguistic approach', *Discourse Studies*, 7(4-5): 585-614.

Dressler, W. and R. Wodak (1982), 'Sociophonological methods in the study of sociolinguistic variation in Viennese German', *Language in Society*, 11: 339-370.

Du Bois, J.W. (2007), 'The stance triangle', in: R. Englebretson (ed.) *Stancetaking in Discourse*. Amsterdam and Philadelphia: John Benjamins. 139-182.

Englebretson, R. (2007), 'Stancetaking in discourse: an introduction', in: R. Englebretson (ed.) *Stancetaking in Discourse*. Amsterdam and Philadelphia: John Benjamins. 1-25.

Fairclough, N. (1992), *Discourse and Social Change*. Cambridge: Polity Press.

Halliday, M.A.K. and C.M.I.M. Matthiessen (2004), *An Introduction to Functional Grammar*, 3rd edition. London: Arnold.

Heal, F. and C. Holmes (1994), *The Gentry in England and Wales, 1500-1700*. Stanford, California: Stanford University Press.

Hyland, K. (2001), 'Humble servants of the discipline? Self-mention in research articles', *English for Specific Purposes*, 20(3): 207-226.

Kress, G. (1989), 'History and language: towards a social account of linguistic change', *Journal of Pragmatics*, 13: 445-466.

Layder, D. (2003) [1997], *Modern Social Theory. Key Debates and New Directions*, 2nd edition. London and New York: Routledge.

Mühlhäusler, P. and R. Harré (1990), *Pronouns and People: The Linguistic Construction of Social and Personal Identity*. Oxford: Basil Blackwell.

Nurmi, A. and M. Palander-Collin (2008), 'Letters as a text type: interaction in writing', in: M. Dossena and I. Tieken-Boon van Ostade (eds.) *Studies in Late Modern English Correspondence. Methodology and Data*. Bern: Peter Lang. 21-49.

Palander-Collin, M. (2006), '(Re)constructing style and language as social interaction through first- and second-person pronouns in Early Modern

English letters', in: I. Taavitsainen, J. Härmä and J. Korhonen (eds.) *Dialogic Language Use*. Helsinki: Société Néophilologique. 339-362.

Palander-Collin, M. (2009a), 'Patterns of interaction: self-mention and addressee inclusion in letters of Nathaniel Bacon and his correspondents', in: A. Nurmi, M. Nevala and M. Palander-Collin (eds.) *The Language of Daily Life in England (1400-1800)*. Amsterdam and Philadelphia: John Benjamins. 53-74.

Palander-Collin, M. (2009b), 'Variation and change in patterns of self-reference in early English correspondence', *Journal of Historical Pragmatics*, 10(2): 234-259.

Porter, R. (1990) [1982], *English Society in the Eighteenth Century*. Harmondsworth, Middlesex: Penguin Books.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Rosenheim, J.M. (1998), *The Emergence of a Ruling Order. English Landed Society 1650-1750*. London and New York: Longman.

Samson, C. (2004), 'Interaction in written economics lectures. The metadiscursive role of person markers', in: K. Aijmer and A.B. Stenström (eds.) *Discourse Patterns in Spoken and Written Corpora*. Amsterdam and Philadelphia: John Benjamins. 199-216.

Scott, M. (2004-2007), *Oxford WordSmith Tools*, version 4.0. Oxford: Oxford University Press.

Seddon, P.R. (2006), 'Holles, John, first Earl of Clare (d. 1637)', in: *Oxford Dictionary of National Biography*. Oxford University Press, September 2004; online edition, May 2006. http://www.oxforddnb.com/view/article /13554, accessed 14 December 2007.

Taavitsainen, I. (2002), 'Historical discourse analysis. Scientific language and changing thought-styles', in: T. Fanego, B. Méndez-Naya and E. Seoane (eds.) *Sounds, Words, Texts and Change*. Amsterdam and Philadelphia: John Benjamins. 201-226.

Wrightson, K. (2003) [1982], *English Society 1580-1680*. London and New York: Routledge.

Sort of and *kind of* in political discourse: Hedge, head of NP or contextualization cue?

Anita Fetzer

Leuphana University Lüneburg

Abstract

This contribution argues for an integrated approach to the analysis of sort of / kind of*, assigning them the status of contextualization cues. It combines quantitative with qualitative, context-sensitive methodologies accounting for distribution, function and co-occurrence. The spoken data comprise ten political speeches (35,844 words) and 22 political interviews (150,044 words), displaying the constructions [DET ADJ* sort of / kind of *NP], [DET* sort of *NP], [QUANTIFIER* sort of / kind of *NP], [*sort of / kind of *VP].*

The local linguistic context of sort of / kind of *contains discourse markers, other hedges, interpersonal markers, question tags and relative clauses.* Sort of / kind of *also co-occur with expressions of epistemic modality realized through modal verbs and modal adverbs, and with cognitive verbs and verbs of communication used as parentheticals. Their most frequent and thus unmarked variant is the more-fuzzy function which co-occurs with other devices expressing indeterminacy. Their marked variant is the less-fuzzy function, which is fronted or co-occurs with relative clauses.*

1. Introduction

Sort of / kind of are used frequently in spoken and written discourse across different registers and genres, with a significantly higher frequency in spoken English (Aijmer 1984; Andersen 2000; Longman Dictionary of Contemporary English (LDCE) 2003). This not only holds for their attested use as qualifier, e.g. adapter (Prince, Frader and Bosk 1982), adjuster (Aijmer 2002), degree word (Bolinger 1972), epistemic modal form (Coates 2003) or hedge (Lakoff 1975), but also for their use as discourse particle, e.g. hesitation marker and discourse connective (Aijmer 2002; Celle and Huart 2007). In their function as lexical head with postdeterminer, they are more frequent in written English (De Smedt, Brems and Davidse 2007: 249).

Regarding their semantics, the type nouns *sort* and *kind* can be defined from a parts-whole and a whole-parts perspective: *kind* refers to "one of the different types of a person or thing that belong to the same group" (LDCE 2003: 887), and *sort* denotes "a group or class of people, things etc. that have similar

qualities or features" (LDCE 2003: 1580). The two type nouns do not denote the thing, as such, to which they refer, but rather a generalized class.

The parts-whole perspective adopts a bottom-up view by singling out representatives which meet almost all of the necessary conditions to count as 'the ideal candidate'; this implies that the representatives are not members par excellence but members allocated only somewhere near the core. The whole-parts conception adopts a top-down view by subsuming all possible representatives which may, in principle, meet the necessary conditions to count as 'the ideal candidate'. The latter represents more of a generalized perspective; it does not intend to account for the allocation of a particular object to a cognitive prototype. Against this background, the type nouns *kind* and *sort* need to be conceived of as a superordinate cognitive prototype with fuzzy boundaries.

As a consequence of their definition as a superordinate cognitive prototype (or a category with fuzzy boundaries), the type nouns *sort* and *kind* contain semantically related subcategories. If *sort/kind* are used to refer to a group of humans for example, possible members need to have the necessary features of 1. +animate, 2. +human, 3. +rational, 4. +intentional and 5. +cultured to classify as a member of that category. If the type noun is the lexical head of a construction qualified by a postmodifier, for instance *a sort/kind of* human being, it refers to a particular subcategory of humans, for instance males, females, young ones, old ones, blue-eyed ones or green-eyed ones.

The superordinate category defines the standard, against which a particular extra-linguistic or discourse-internal object is evaluated. If it meets all of the necessary features, it is allocated to the core of the cognitive prototype and assigned the status of a representative par excellence. If it does not meet them fully, it cannot be allocated to the core but rather needs to be placed somewhere between core and periphery as an almost *kind of / sort of* best example.

The allocation of an object closer to the periphery of the cognitive prototype assigns it the status of a particularized member. That status is conceived of along the lines of the differentiation between a generalized and particularized conversational implicature (Grice 1975; Levinson 1983, 2000), and depends strongly on linguistic and social context. While the generalized scenario refers to the lexical head+postmodifier construction expressing dual reference to the type nouns *sort/kind* and to the postmodifier, the particularized scenario refers to the lexical head+postmodifier construction expressing single reference to the particular object. Here the type noun is assigned an attributive status (Davidse, Brems and De Smedt 2008).

In the example discussed above, *a sort of human being* can refer to some particularized animate creature, e.g. a pet, which does not fully qualify as a human being as it lacks civilization, rationality or intentionality, for instance. Here, the attribution of the creature at hand to the superordinate category of human being expresses additionally the speaker's evaluative stance, which tends to be ironic, and is classified as subjectification in the paradigm of grammaticalization (Traugott 1995). It needs to be pointed out, however, that the construction, as such, does not prefer either interpretation, and that the interpretation as

realizing single or dual reference depends strongly on context. In the former scenario, subcategories are attributed to their entailing superordinate category, and in the latter scenario a member qualifying partially for inclusion is allocated somewhere close to the periphery of the cognitive prototype.

The classification of the construction *sort of / kind of* as qualifier or as head of NP depends strongly on context, viz. social context as regards the speaker's attitude and evaluation, and linguistic context (or co-text) with respect to co-occurrence patterns, pragmatic enrichment and disambiguation. Against this background, a felicitous examination of *kind of / sort of* requires a discursive framework, which may accommodate the connectedness between participants, and the connectedness between participants, text and context.

In the following, the form, function and scope of *sort of* and *kind of* are investigated in the context of spoken political discourse, accounting for their particularities in the dialogical genre of interview and the monological genre of speech. Particular attention is given to the questions of (i) whether *sort of / kind of* fulfil similar functions in discourse as head of NP and as the adverbial-based category of hedge, and (ii) whether a discursive frame requires a different classification, namely one based on contextualization cue (Gumperz 1977, 1991). The methodological framework employed is an integrated one. It combines quantitatively oriented corpus analysis with qualitatively oriented sociopragmatics and cognitive pragmatics, supplementing distribution and co-occurrence with prototypicality, intentionality of communicative action and genre (Fetzer 2004).

The next section examines the theory and practice of *sort of / kind of*. Section 3 introduces the data; section 4 presents a genre-specific analysis accounting for distribution, collocation and scope. Section 5 examines their functions and section 6 summarizes the results obtained and concludes.

2. *Sort of* / *kind of*: theory and practice

Sort of / kind of serve multiple functions in discourse: they can be head of NP, modifier with qualifying function and discourse particle. As a consequence of the multilayered form-function mapping, their communicative meaning depends strongly on context. In the following, the different functions of *sort of / kind of* are analysed and exemplified. The examples stem from political-discourse data,[1] which are introduced in detail in the next section. Functions and categories reported in current research, which do not occur in the political data, are adopted accordingly.

2.1 *Sort of* / *kind of* as hedge

The most prominent function of *sort of / kind of* is that of a modifier with qualifying function, viz. hedge. In synchronic grammars of English, hedges are allocated to the word class of adverb and to the grammatical function of adverbial.

In the *Grammar of Contemporary English* (Quirk et al. 1985), hedges are part of the category of subjunct with narrow scope, which is defined as having "a subordinate [...] role in comparison with adjuncts; they lack the grammatical parity with other sentence elements" (Greenbaum and Quirk 1990: 176). In the heterogeneous class of subjuncts, hedges belong to the subcategory of intensifier as downtoner with compromiser function. Compromisers have only a slight lowering effect. They refer to the appropriateness conditions of an utterance signalling that they "reach out towards an assumed norm but at the same time reduce the force of the verb" (Quirk et al. 1985: 455).

In the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), hedges belong to stance adverbs, which are defined as "adverbials that overtly mark a speaker's or writer's attitude to a clause or comment about its content. They can be divided into three categories: epistemic, attitude and style" (Biber et al. 1999: 382). Hedges are defined as signifying imprecision. They tend to focus on a particular element in the clause, suggesting that the proposition (or part of it) is imprecise.

In his examination of adverbs and adverbials, Ungerer (1988) differentiates between propositional adverbials and scope adverbials. The former are a constitutive part of the proposition and for this reason truth-conditional, while the latter are metalinguistic. In his frame of reference, English is assigned the status of a scope-initial language, taking linguistic context explicitly into consideration. Scope is given a semantic interpretation. It manifests itself through a carrier of scope, and carriers of scope are hierarchically organized along the following cline: sentence mode > tense > modals > not-negation > scope adverbials. This means that *sort of / kind of* can have both narrow scope, qualifying a single constituent, and wide scope, qualifying more than one constituent.

In cognitive linguistics, hedges are defined in the framework of prototypicality conditions, or to use Lakoff's words in his original definition of hedges, as "words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy" (Lakoff 1975: 234). Against this background, hedges can be differentiated into more-fuzzy hedges, making things fuzzier, and less-fuzzy hedges, making things less fuzzy (Fetzer 1994, 2004): the former anchor the object, constituent or clause, over which they have scope, closer to the periphery of a cognitive prototype, while the latter anchor it closer to its core. That dual function is of great importance to assigning *sort of / kind of* the status of contextualization cue and will be elaborated on in section 2.4. The more-fuzzy function can be explicated with the metalinguistic comment 'what you might call' (Kay 1984), indicating 'loose interpretation' (Andersen 2000), and the less-fuzzy function can be explicated by the metalinguistic comment 'what is more precise'.

In her analysis of *sort of*, Aijmer (2002) pays particular attention to its multiple functions on the interpersonal and interactional domains of discourse. Her research corroborates the results obtained in synchronic grammar with respect to the expression of evidential meaning, in particular imprecision. For this reason, the use of *sort of* may affect the truth value of a proposition. On the interpersonal plane of communication, *sort of* is assigned the status of an

affective marker hedging strong opinions by toning down the impact of a communicative contribution, thus reducing its imposition and at the same time establishing common ground. This is also reflected in Brown and Levinson's research on politeness (Brown and Levinson 1987), in which hedges refer indexically to one or more of the Gricean maxims (Grice 1975), triggering a conversational implicature in order to signify interpersonal meaning, for instance positive politeness (or solidarity) and negative politeness (or respect). On the interactional plane of discourse, *sort of* may signify self-repair.

In De Smedt, Brems and Davidse's examination of type nouns (De Smedt, Brems and Davidse 2007), the hedging function is further refined with respect to the use as attributive modifier, e.g. *a straightforward kind of guy* [[DET] [[ADJ *sort of*] [N]]] (IR 2001b), in which the attribute comprises both ADJ and *kind of*. Furthermore, they identify the semi-suffix use of *kind of / sort of* in the collocation *superhero-kind-of costume*, in which the type-noun string is interpreted as a whole. Here, the classifier *superhero* is hedged, signifying the speaker/writer's implicit evaluation. That particular use has not been found in the political-discourse data and seems to occur primarily in contexts in which ad-hoc meaning is constructed (Mauranen 2004).

In the following, the function of *sort of / kind of* is examined more closely. In extracts[2] (1) and (2), they are assigned the function of signalling that the lexical item or items, over which they have scope, are not in full accordance with the appropriateness conditions of the contribution. They are not met fully, but only to the extent of 'what you might call':

(1) *Jonathan* can we now just stop this. *I mean* this is just carrying the thing on and on […]. *And quite frankly*, there is a sort of feeding process you only have to change a comma [IE 1990a]
(2) *Well Jonathan let's* sort of consider the enormity of what you've just said [IE 1990a]

Following Kay, the communicative meaning of *sort of* can be made explicit by inserting the metalinguistic comment 'what you might call' with the status of a parenthetical. That is, (1) would be: "there is a, *what-you-might-call*, feeding process" and (2) would be "let's, *what-you-might-call*, consider the enormity of what you've just said". The hedge has scope over the noun "feeding process" in (1) and the verb "consider" and its argument "enormity" in (2). In both cases, *sort of* indexes the appropriateness conditions, namely those of the noun "feeding process", which is mapped across different cognitive domains and used metaphorically, and those of the verb "consider", which denotes a neutral activity but is used here to refer to an emotionally loaded context signified by the verb's argument "enormity".

In both excerpts, the linguistic context displays further attitudinal and discourse-structuring devices, and in both cases the interviewee addresses the interviewer with his first name, summoning him to return to a neutral style of interviewing. There are other attitudinal and discourse-structuring devices, viz.

the cognitive-verb-based parenthetical "I mean" used to introduce a self-reformulation, the connective "and" and the style disjunct "quite frankly", which signals forthcoming criticism. In (2) the negative context is indicated by the negative discourse marker "well", while the forthcoming imposition is attenuated by the marker of solidarity "let's".

Dyadic discourse often displays a fine-grained interplay between discourse-organizing devices and attitudinal markers, boosting and attenuating the degree of speaker commitment, often within a single contribution (Fetzer 2008).

Not meeting all of the necessary features to be attributed to the core of the cognitive prototype, which may refer to an extra-linguistic object, to a discourse-internal entity or to a communicative contribution, is not only a constitutive part of the use of *sort of / kind of* as hedge but also of its use as discourse particle, as is examined below.

2.2 *Sort of / kind of* as discourse particle

The prime function of hedges, and in particular of *sort of / kind of*, is interpersonal and interactional (Aijmer 2002; Brown and Levinson 1987; Mauranen 2004). This is also corroborated by Le Lan's research on connectives, a cover term referring to discourse markers, hedges, comment clauses and particles (Le Lan 2007). All of the devices are indexical, dynamic and multifunctional. Le Lan assigns them a cognitive function, viz. they signify some mental movement from the speaker's to the hearer's point of view, and vice versa, regulating and organizing the fundamental triad of speaker, hearer and discourse. She draws the conclusion that their function of indicating intersubjective positioning makes them metacommunicative rather than metalinguistic: "Hedges enable the speaker to qualify the illocutionary force of his/her utterance by making things clearer or more obscure" (Le Lan 2007: 109). Thus, their primary goal lies in the regulation and organization of discourse. So, what is particular about the function of *sort of / kind of* as discourse particle and in what way does the particle function differ from the hedge function?

In the analysis of *sort of / kind of* as hedge, their relational meaning has been stressed, relating a lexical expression or a group of lexical expressions to an extra-linguistic or discourse-internal entity. The qualification of those relations as not meeting all of the necessary features to classify as the 'object proper' is a constitutive part of the qualifier meaning. As discourse particle, *sort of / kind of* no longer primarily signify the core-peripheral relationship between extra-linguistic or discourse-internal object and cognitive prototype. Their meaning is procedural only, indexing interactional and interpersonal domains, assigning them some degree of fuzziness, thus allowing for intersubjective manoeuvring and positioning. This does not, however, generally cause communicative problems as the speaker assumes that "the hearer will be able [to] figure out the meaning of what is said even if it [is] only approximate" (Aijmer 2002: 209).

In the political-discourse data, there was only one instance in which the use of *sort of / kind of* counts as a discourse particle, indicating that the speaker is aware that his contribution may not be in full accordance with the appropriateness

conditions. The possible mismatch may be due to the social-context constraint of political discourse having become a subcategory of media discourse and of professional discourse (Fetzer 2008), which puts a politician under the obligation to present unambiguous and clear statements, and to provide relevant information. Against this background, providing approximate answers with informational gaps, which need to be filled by the hearer, is not necessarily considered to count as a professional, but rather as an inappropriate, response.

Excerpt (3) stems from a political speech delivered by Tony Blair. The discourse particle *kind of* occurs at a stage in discourse where the audience applauds, expressing agreement with the politician. To adhere to the modesty maxim (Leech 1983) and to save his face, Blair reacts with the face-saving strategy of toning down the impact of praise, which is ratified by the audience with laughter:

(3) The city of the Olympic Games for Britain in 2012. Let me tell you what won that bid. Yes, we had a magnificent team led by Seb Coe, a great London Mayor who backed it to the hilt, a country behind us [*applause*] kind of gets better every year that line actually [*laughter*] but what won it was London itself [POL 2005f]

The discourse particle *kind of* occurs right after the audience's applause, which counts as an interruption in the genre of political speech. It allows the politician to retain the floor. At that stage in discourse, the discourse particle does not express vagueness with respect to the verbal phrase "gets better" nor does it denote a core-periphery relation. Rather, it indexes the interactional domain of discourse, signifying responsive acknowledgement. Simultaneously, it indexes the interpersonal domain, toning down the impact of the applause while introducing a side sequence realized through the metalinguistic comment "gets better every year that line" signifying self-irony. In De Smedt, Brems and Davidse's frame of reference, that metalinguistic use would qualify as a marker of quoted thought (De Smedt, Brems and Davidse 2007: 248).

In mundane everyday talk, the use of *kind of / sort of* as a discourse particle is far more frequent than in political discourse, which is an instance of institutional discourse. The particles fulfil an important role in the negotiation of interpersonal and interactional roles in everyday talk, especially in dyadic or multi-party informal conversation (Aijmer 2002; Davidse, Brems and De Smeldt 2008). In the context of institutionalized media talk, the interpersonal and interactional domains are pre-structured and underlie particularized constraints, such as neutralism (Fetzer 2000; Greatbatch 1998).

In the following, the use of *sort of / kind of* as lexical head is examined, which is more frequent in the political-discourse data.

2.3 *Sort of* / *kind of* as lexical head

In his examination of *sort of*, Andersen (2000) differentiates between two different meanings of *sort of*: *sort of*$_1$ is a modifier signalling loose interpretation

of a vague and less well defined entity, and *sort of*$_2$ is analysed as head of a noun phrase with postmodifier and seen as equivalent to 'type of' indicating category membership (cf. Andersen 2000: 48). In spite of the different grammatical functions, *sort of*$_1$ and *sort of*$_2$ share the common feature of addressing the nature of the connectedness between speaker, extra-linguistic or discourse-internal object and the lexical item selected to refer to that 'object'. Both indicate that the lexical item selected to refer to the object at hand does not realize the connectedness of a perfect match. Had it been a perfect match, the object at hand would have been referred to by the postmodifier only, and not by the lexical head *sort/kind* (=N_1) and postmodifier N_2. Hence, attributing an 'object' to a category, classifying it as a member of that category and naming it accordingly is functionally equivalent to assigning it the status of best example. For instance, referring to a living entity, which has the features 1. +animate, 2. +human, 3. +rational, 4. +intentional and 5. +cultured, as a 'human being' is functionally equivalent to classifying it as a human being par excellence. Should a speaker/writer refer to that living entity as *a sort of human being*, s/he evaluates the connectedness between extra-linguistic/discourse-internal object and lexical items selected to refer to that object as not that of a perfect match. Instead, s/he allocates it somewhere between the core of the cognitive prototype and its periphery, thus assigning it loose interpretation. In a similar vein, *a sort of interview* is not synonymous with *an interview*. Against this background, *sort of*$_1$ and *sort of*$_2$ may be assigned different grammatical functions in discourse, but their communicative functions share the common feature of signifying fuzziness.

The collocations *sort of / kind of* in their function as lexical head (=N_1) with postmodifier (=N_2) refer to a presupposed general category of an object (=N_2), while at the same time expressing the speaker/writer's evaluation that a particular N_2 cannot be assigned the status of best example but needs to be interpreted loosely. This interpretation is in discordance with its classification as a binominal configuration with status-identical N_1 and N_2. However, it is in accordance with De Smedt, Brems and Davidse's claim that the whole NP designates a subclass (De Smedt, Brems and Davidse 2007).

In cognitive-semantic terms, the object at hand cannot be attributed to the core of a cognitive prototype as it does not fully share its constitutive features. This is the case in (4), where the type noun *kind* singles out a particular subcategory of "reforms", namely the ones that are part of the politician's (but not of another politician's) programme; and it is also the case in (5) where the type noun *kind* singles out a particular category of "people", namely the ones not broken by terror, again implicitly contrasting them with the category of people who are broken by terror. Of further relevance to the examination of the function of *kind of / sort of* is their co-occurrence with the subjectivizing device "in your view", referring to the interviewee's personal opinion in (4), and with the first-person plural self-reference in (5), also expressing subjectivity:

(4) I understand it therefore that *in your view* it is extremely important that the kind of <u>reforms</u> *that you have advocated are put in place*? [IR 1990a]

(5) We are*n't* the kind of people *to be broken by terror*. We won't appease it. And together, as one nation, we will confront it [POL 2005b]

From a pragmatic perspective, the category-subcategory relation not only implicates a container-contained relationship but also one indicating contrast. By singling out a particular subcategory and specifying it with a definite determiner and with a defining relative clause supplying relevant contextual information, viz. the people not broken by terror or the reforms advocated by the politician, all of the other possible subcategories are excluded. Hence, the *DET sort of / kind of N* construction signals an upcoming discourse topic, corroborating De Smedt, Brems and Davidse's (2007) context-based analysis.

In the following, *sort of / kind of* are examined in a discursive frame of reference, in which function and co-occurrence are of key importance.

2.4 *Sort of / kind of* as contextualization cue

The collocations *sort of / kind of* have been defined as multiple form-function mappings. To assign the forms one of their possible functions, viz. hedge, discourse particle or head of NP, the role of context and the co-occurrence of *sort of / kind of* with other linguistic devices expressing categorization, determinate meaning, subjectification and fuzziness are of great importance. Additionally, their function in realizing anaphoric and cataphoric reference in discourse needs to be taken into consideration, as *sort of / kind of* can construe anaphoric and cataphoric generalization relations (De Smedt, Brems and Davidse 2007). The function of realizing anaphoric and cataphoric reference is a general feature of degree words, which may both identify and intensify (Bolinger 1972).

In discourse, hedges and type nouns target the act of reference by contextualizing it and by particularizing its conditions of use. Hedges and type nouns signify the contextual conditions under which an act of reference is felicitous thus making explicit the connectedness between extra-linguistic object, cognitive prototype and linguistic expression. For this reason, they convey relational meaning, relating extra-linguistic objects, linguistic expressions and cognitive prototypes. Hedges and type nouns indicate that the lexical expression which they make fuzzier or which they type is not the best example of the prototype at hand but only *some type/sort/kind of* a member. Hence, type nouns and hedges make manifest an instruction-of-interpretation about how the speaker speaker-intends the nature of the connectedness between extra-linguistic object, cognitive prototype and linguistic expression, and how s/he intends the hearer to interpret that connectedness. A linguistic device, which is relational by definition and accounts explicitly for the connectedness between form, function and context, is that of a contextualization cue adopted from the research paradigm of interactional sociolinguistics.

Interactional sociolinguistics provides a truly dialectical frame of reference accommodating the examination of parts, such as objects, constituents and clauses, and the examination of wholes, that is discourse. Its unit of investigation is the speech activity[3] in which language is assigned the status of a socially

situated form anchored to the intentionality and indexicality of communicative action. Hence, linguistic variation and alternation are not random or arbitrary, but communicatively functional and meaningful. This is reflected in the functional category of contextualization cue, which "serve[s] to retrieve the contextual presuppositions conversationalists rely on making sense of what they see and hear in interactive encounters" (Prevignano and di Luzio 2003: 9). In order to be able to retrieve contextual presuppositions, inference is given a context-dependent interpretation and assigned the status of conversational inference, connecting logical reasoning with the sociocultural activity of conversation (Gumperz 1991, 2003).

In its original definition, contextualization cue is defined as "any aspect of *surface* form of utterances, which when mapped onto message content, can be shown to be *functional* in the *signaling of interpretation frames*" (Gumperz 1977: 199). Not only is contextualization cue a metalinguistic concept but it is also metacommunicative, signalling procedural meaning. In more recent examinations its relational nature is highlighted, connecting it with cognitive linguistics on the one hand, and functional linguistics on the other:

> They serve to highlight, foreground or make salient certain phonological or lexical strings *vis-à-vis* other similar units, that is they function relationally and cannot be assigned context-independent, stable, core lexical meanings. Foregrounding processes, moreover, do not rest on any one single cue. Rather, assessments depend on cooccurrence judgments [...] that simultaneously evaluate a variety of different cues. When interpreted with reference to lexical and grammatical knowledge, structural position within a clause or sequential location within a stretch of discourse, foregrounding becomes an input to implicatures, yielding situated interpretations. (Gumperz 1991: 232)

In the analysis of the theory and practice of *sort of* and *kind of*, the expressions have been defined as expressing relational meaning, relating extra-linguistic objects, cognitive prototypes and linguistic expressions. Assigning these devices the status of contextualization cue has the consequence that the linguistic forms *sort of / kind of* cannot, per se, fulfil the functions of making things fuzzier or less fuzzy, of typing objects or of expressing procedural meaning. Rather, their local linguistic context and collocations need to be interpreted "*vis-à-vis* other similar units" over which they have scope. Furthermore, the nature of the connectedness between *sort of / kind of* and their collocates needs to be evaluated because the "assessments depend on cooccurrence judgments [...] that simultaneously evaluate a variety of different cues". Only then can they instruct the hearer about possible speaker intentions, that is whether the communicative contribution (or one or more of its constitutive parts) is intended to be allocated closer to the core of the prototype or prototypical scenario as less-fuzzy, or whether it is intended to be allocated closer to the periphery as more-fuzzy.

The research tradition of hedges has focussed primarily on the more-fuzzy function, refining the original definition by accommodating more fine-grained categories. Most of the investigations have not, however, explicitly accommodated the fact that fuzziness is a scalar concept, anchoring 'the object', over which a hedge has scope, to a singled out area on a scale. By referring to a particular area and by anchoring the object at hand to that area, the object is bounded by more fuzziness on the one side of the cline and by less fuzziness on the other side of the cline. Looked upon from a reasoning and accessibility-of-information perspective, for something to be classified as more-fuzzy, for example for an interview to be classified as *a sort of interview* or for asking questions to be classified as *sort of asking questions*, the interview or the activity of asking questions need to have been contrasted with their less-fuzzy counterparts, that is *an interview par excellence* or *asking questions as it is usually done.* The anchoring of an object to a particular area on a scale is not arbitrary but depends on the semantics of the qualifying expression(s). The higher degree of indeterminateness, for instance *some sort of interview* or *something like sort of asking questions*, the closer the object is anchored to the periphery. The higher degree of determinateness, for example *this sort of interview* or *what sort of asking questions*, the closer the object is anchored to the core.

In the remaining part of the chapter, the distribution, collocation, scope and function of *sort of* and *kind of* are examined within the social context of political discourse, but first the data are presented.

3. The data

The dyadic data comprise 22 full-length dyadic political interviews, as systematized in table 1a in the appendix. The 1990-data were recorded from the programme *On the Record* (OTR) on BBC1; the 1997 and 2001 data are pre-election interviews (PEI) with the leaders of the leading British parties. The 2003 interview is about the war in Iraq. The data contain 150,044 words with 52 tokens of *sort of / kind of*, which is roughly equivalent to 3.46 instances per 10,000 words. *Sort of / kind of* are significantly less frequent than in the *London-Lund Corpus* section of informal conversation which has roughly 39 instances per 10,000 words (Aijmer 1984).

The monologic data comprise ten political speeches delivered by leading British politicians at the Conservative, Labour and Liberal Party Conferences, as systematized in table 2a in the appendix. They contain 35,844 words with twelve tokens of *sort of / kind of*, which is roughly equivalent to 3.34 instances per 10,000 words.

Sort of / kind of is far more frequent in informal everyday conversation than in political discourse. This is primarily due to the social-context constraints of political discourse having become an instance of media discourse and of professional discourse. In institutional dyadic discourse, the turn-taking mechanism is context-sensitive. There is no self-selection and, as a consequence, less negotiation of interactional rights and obligations. In institutional monologic

discourse, that type of negotiation of meaning is superfluous. Furthermore, the expression of fuzziness within the domain of political discourse as professional discourse transmitted through the media is also a context-sensitive endeavour as political agents, viz. journalists as interviewer and politicians as interviewee or public speaker, are expected to provide relevant and clear information. For that task, too much fuzziness is not appropriate.

In her analysis of the *London-Lund Corpus*, Aijmer (1984) found a preference for *sort of* over *kind of*. In the political-discourse data, that is no longer the case.

In the following, the results of the micro-analysis of political discourse are refined, accommodating distribution, collocation and function.

4. Distribution and collocation of *sort of / kind of* in the political interviews and speeches

In the data at hand, there are NP- and VP-anchored collocations. The former show variation with respect to determiner, but not so much with respect to quantifier:

[$_{NP}$ DET (ADJ) sort / kind of N]

(6) ...the sort of spending *you're planning with the health service*… [IR 2001f]
(7) …the kind of reforms *that you have advocated*… [IR 1990a]
(8) …when I left the government I was not for a moment of the view that there was a sort of return ticket and I haven't changed my mind about that…[IE 1990a]
(9) …quite frankly there is a kind of feeding process… [IE 1990a]
(10) …this sort of *erm* detour from reality… [IE 1990a]
(11) …are associated with this kind of situation… [IR 1990d]
(12) …that can happen in that sort of situation… [IE 2003]
(13) …and that kind of tragedy *that happened*… [IE 2001d]
(14) …what kind of complacency is that… [POL 2005d]
(15) …create another sort of media story… [IE 1990a]
(16) …you're a straightforward kind of guy… [IR 2001b]

The [$_{NP}$ DET (ADJ) sort / kind of N] construction is also referred to as binominal construction, and its internal configuration is examined thoroughly by Davidse, Brems and De Smedt (2008) and De Smedt, Brems and Davidse (2007). The goal of this section does not lie in the internal configuration of the construction, but rather in its function in discourse. The diverse determiners (definite, indefinite, demonstrative and interrogative) co-occur with both type nouns. In the data, the numerative "another" only collocates with the type noun *sort* (example 15), and the attributive-adjective configuration in (16) only collocates with the type noun *kind* [$_{NP}$ DET ADJ kind of N], creating the ad-hoc classifying class of straight-

forward persons, which implicates the contrastive set of non-straight-forward persons.

[$_{NP}$ QUANTIFIER sort / kind of N]

(17) …of setting up some kind of task force… [IR 1990c]
(18) …that you will get some sort of story because… [IE 1990a]

In the data, *kind of / sort of* only collocate with the quantifier 'some'. Again, this seems to be a genre-specific constraint. In informal conversation and in written data, there is more variation with respect to quantifier collocation, as is demonstrated by Davidse, Brems and De Smedt (2008).

[$_{VP}$ sort / kind of V]

(19) …kind of gets better every year… [POL 2005f]
(20) …we're more sort of spun against than spinning… [IE 2001a]

In the data, the VP collocates with *kind of* and *sort of*, exemplifying the prevailing definition of a hedge. Its function is to index the appropriateness conditions, signifying fuzziness or approximation (Aijmer 2002).

The following tables systematize the *sort of / kind of* anchored collocations in the political interviews and political speeches. While there is quite some variation in the interviews, the speeches display only few variants. The numbers in brackets refer to the absolute frequency of the collocations with an almost even distribution across the type nouns *sort* and *kind* in the NP-based configurations. With the VP, only the type noun *sort* collocates. Table 1 presents the interview-based results:

Table 1: Distribution of *sort of / kind of* in the interviews

a sort of NP **(7)**	a kind of NP **(6)**
the sort of NP **(5)**	the kind of NP **(9)**
sort of NPs (1)	
this sort of NP (2)	this kind of NP (2)
that sort of NP (2)	that kind of NP (2)
another sort of NP (1)	
what sort of NP (2)	
	any kind of NP (1)
some sort of NP (2)	some kind of NP (2)
sort of VP **(9)**	

In the political speeches, the attributive modifier use anchored to the interrogative determiner *what* collocating with the type noun *kind* is the most frequent configuration. According to De Smedt, Brems and Davidse (2007: 234), the

attribute, viz. the *wh*-word invokes "a quality of specific feature". In the LDCE (2003: 1580) the collocation *what sort of* is described as conveying the speaker's negative attitude, in particular being angry about something. The type-noun anchored collocations are systematized in table 2:

Table 2: Distribution of *sort of / kind of* in the speeches

	kind of VP (1)
what sort of NP **(1)**	what kind of NP **(5)**
	the kind of NP (2)
this sort of NP (1)	
that sort of NP (1)	that kind of NP (1)

The distribution of the type nouns' collocates show preferences for more determinate (or less-fuzzy) collocates in the monologic genre of speech and a more balanced distribution of more determinate and more indeterminate collocates in the dialogic genre of interview. To analyse the function of *sort of / kind of* in the political discourse, their linguistic context needs to be taken into consideration more closely.

5. Function of *sort of / kind of* in the political interviews and speeches

Lexical expressions can be examined with respect to their functions in a clause or sentence, viz. head of NP, modifier or discourse particle; with respect to their semantics, viz. type noun, qualifier or discourse-organizing device, as has been done in section 2; and they can be examined with respect to their pragmatic function in discourse, as has already surfaced in the analysis of contextualization cues and which is elaborated on in the following.

A pragmatic frame of reference is based on the premise that participants, viz. speakers and hearers, perform communicative actions in context (Fetzer 2004). To account for the inherent connectedness between participants, language, language use and context, participants are conceived of as rational and intentional agents who perform their communicative acts accordingly. These necessary conditions are required for participants to act purposefully, to decode literal meaning and to infer meaning which goes beyond the level of what has been said (or written). Devices serving this purpose are indexical expressions (Bar-Hillel 1998), such as illocutionary force indicating devices in the framework of speech act theory (Searle 1969) or contextualization cues in interactional sociolinguistics. Indexical expressions are defined from a relational perspective, based on communicative intention and co-occurrence.

To account for the impact of co-occurrence and co-occurrence patterns on the function of *sort of / kind of*, their linguistic context is taken into consideration more closely. In the investigation of the data in the previous section, co-occurrence patterns have already surfaced. In the political interviews *sort of / kind of* co-occur with discourse markers, signifying particularized contexts. For

instance, *well* indicates a negative context (Schiffrin 1987), *now* and *look* are used to attract the hearer's attention, signalling contexts of immediate relevance, and *first* and *second* signify an argumentative sequence. Hedges in their prevailing interpretation as more-fuzzy hedges, such as *more or less*, *in the sense*, *somehow* or *almost* signify vague contexts, which are open for a negotiation of meaning (Andersen 2000).

In the interviews, *sort of / kind of* not only co-occur with the devices listed above, but also with the communicative strategy of understatement and its constitutive part of double negation (Brown and Levinson 1987). They co-occur with interpersonal markers, for instance *you know*, *if you like* or *to my best knowledge*, and with question tags. Furthermore, their local linguistic context displays expressions of epistemic modality realized by the modal verbs *can*, *must*, *shall* and *will*, the modal adverbs *perhaps*, *possibly*, *probably* and *obviously*, a number of cognitive verbs used parenthetically and by verbs of communication, such as *I mean*, *I know*, *I think*, *I believe*, *it seems*, *I feel* and *talk* or *suggest* signifying subjectification and intersubjectivity (Verhagen 2005).

In the previous section, *sort of / kind of* have been assigned the status of contextualization cue foregrounding salient information. They indicate whether the speaker intends her/his communicative contribution, or one of its constituents, to count as a member allocated closer to the core of a cognitive prototype, viz. a more prototypical case of representation, or whether s/he intends it to be allocated closer to the periphery, counting as a less prototypical case of representation. While the former makes it a less fuzzy case of representation, the latter assigns it the status of more fuzzy.

5.1 More-fuzzy contextualization cue

As type nouns, *sort* and *kind* refer to a general superordinate category, but never to a particular 'object' or 'scenario' in context. For this reason, they are intrinsically indeterminate. From a semantic perspective, they are more-fuzzy, as has been corroborated by research on their function in everyday conversation, where *sort / kind* co-occur with general nouns and general verbs (Aijmer 1984). Against this background, the more-fuzzy interpretation of *sort of / kind of* is seen as the unmarked variant (or default).

However, *sort of / kind of* have undergone a process of grammaticalization (Davidse, Brems and De Smedt 2008), and their linguistic form has been assigned a number of different functions in discourse. For this reason, the local context generally displays further cues, which support a more-fuzzy or a less-fuzzy interpretation. So, what other devices does the local context of *sort of / kind of* contain in the political-discourse data?

In the extracts (21), (22) and (23) *sort of / kind of* collocate with a VP, over which they have scope. This is the standard collocation for *sort of / kind of* to be assigned the status of qualifier (e.g. hedge, adapter) in synchronic English grammar (e.g. Biber et al. 1999; Quirk et al. 1985). In spite of this unambiguous configuration, the linguistic context of all of the extracts contains linguistic devices, viz. "more or less", "the other day", "sometimes" in (21), which express

vagueness. In the IR's contribution, the expressions of vagueness co-occur with the marker of common ground "you know" indexing the interpersonal domain of communication. In the IE's contribution, the expressions of vagueness co-occur with the negative discourse marker "well" and the reformulation device "in the sense" making the second part of the contribution less fuzzy:

(21) *More or less*, yes. On on on spin, *you know*, Peter Mandelson said *the other day*[…]. Do you accept that there's been too much spinning? [IR 2001a]
Well, I I sometimes think we're more sort of spun against than spinning, *in the sense* that people have this huge thing about it… [IE 2001a]

There is a similar pattern in extracts (22) and (23), yet the implicated meaning is different. In (22), there is the quantifier "some", which is indeterminate regarding its semantics. From a discursive perspective, however, "some" collocates with "carefully worked out", particularizing the indeterminateness postulated, while at the same time expressing a negative evaluation of the IR's strategy. *Sort of* with scope over the VP "change the emphasis" is assigned the function of expressing some degree of fuzziness:

(22) and I'm not prepared to be taken down *some road* that you've carefully worked out in order to try and sort of change the emphasis [IE 1990a]
(23) But so when he says there will be no further cuts, Michael Portillo says there will be f f further cuts, he is *erm* sort of whistling in the wind? [IR 2001d]

The situation is slightly different in extract (23), whose linguistic context displays further devices signifying indeterminacy. *Sort of* has scope over the VP "whistling in the wind", attributing some degree of fuzziness to it. The local linguistic context displays the hesitation device *erm* realized right before the hedge, thus supporting the more-fuzzy interpretation. Hesitation indicating approximation is also reflected in an instance of dysfluency connected with the realization of "further". In (22), the IE expresses his negative evaluation of the IR, and in (23) the IR expresses an ironic stance towards both IE and the source of the represented discourse.

For *sort of / kind of* to be assigned the status of a more-fuzzy contextualization cue, the linguistic context needs to contain devices whose semantics are intrinsically indeterminate. If there are no such devices, but intrinsically determinate expressions, *sort of / kind of* assign fuzziness to the constituent(s), over which they have scope, while at the same time expressing the speaker's emotive stance towards communicative contribution and/or co-participant.

In the following, the linguistic contexts in which *sort of / kind of* is assigned a less-fuzzy function are examined.

5.2 Less-fuzzy contextualization cue

In the data analysed, *sort of / kind of* not only co-occur with devices which are semantically indeterminate but also with constructions which assign them a prominent status, such as syntactic fronting, and with defining relative clauses, which supply relevant contextual information making the type noun's postmodifier, which tends to be realized by general nouns, more specific. It needs to be pointed out, however, that in this particular discursive context, general nouns are used with anaphoric reference and therefore are pragmatically enriched and more determinate. Further devices, which make the lexical string more determinate and less fuzzy are demonstratives and definite determiners. Because of the semantics of *sort of / kind of*, their co-occurrence with less-fuzzy devices is assigned the status of a marked configuration (or a non-default).

In extracts (24) and (25) from the interview data, *sort of* collocates with a demonstrative determiner and with the general noun "thing" as its postmodifier both realizing anaphoric reference, thus expressing determinate meaning. This is supported by the politician's personal stance towards politics, the exclamatives "heaven forfend" and "thank goodness", and the reformulation of his personal opinion introduced by "I mean", subjectifying the less-fuzzy information contained in the communicative contribution. In (25) *sort of* collocates with the quantifier "any" and with the postmodifier "referendum" qualified by the adjective "informed" and another qualifier attached to the postmodifier "on the subject", making it more specific and thus less fuzzy:

(24) *Well I think* that politics is a tough trade *I mean heaven forfend* that *erm* exchanges should get personal that sort of thing *thank goodness* has never happened to me. But of course... [IE 1997a]

(25) is there *any chance whatsoever* of happening to form an intelligent debate prior to any kind of informed referendum *on the subject*, because what you've had for the past few years is nothing but nonsense… [AM 2001e]

(26) The government says 'but we all know what we are talking about.' *What* kind of complacency is that? That is no way to make laws… [POL 2005d]

(27) Because *just what* kind of democracy was it that delivered back in May? A democracy which returns an outright majority government on little more than… [POL 2005d]

Extracts (26) and (27) stem from the monologic data, where the type nouns *sort/kind* collocate frequently with the interrogative determiner "what", which may be intensified with the emphasizer "just" as in (27). In both cases, the syntactic structure, viz. the interrogative determiner collocating with type noun and auxiliary, indicates an interrogative configuration, from a formal perspective. Its communicative status, however, is that of a rhetorical question expressing the speaker's stance towards a certain state of affair, i.e. his negative evaluation of the entity referred to in the postmodifier. Thus, instead of providing an open slot

indexed by the interrogative determiner to be filled by the hearer, the speaker indexically refers to a particularized and thus less-fuzzy concept.

For *sort of / kind of* to be assigned the status of a less-fuzzy contextualization cue, the linguistic context needs to contain devices whose semantics are intrinsically determinate, viz. defining relative clauses, qualified postmodifiers or definite, demonstrative and interrogative determiners. Furthermore, the determiners and postmodifiers need to realize anaphoric or cataphoric reference. Analogously to its more-fuzzy counterpart, *sort of / kind of* express the speaker's emotive stance towards communicative contribution and/or co-participant.

In the following, the fine-grained interplay between more-fuzzy and less-fuzzy contextualization cues within a communicative contribution is examined more closely.

5.3 More-fuzzy and less-fuzzy contextualization cues in context

The political-discourse data under investigation are all adopted from present-day televised discourse in the British context. Because of social-context and media-context constraints, in particular the constraint of neutralism (Greatbatch 1998), political discourse is produced and interpreted accordingly (Fetzer 2000, 2006).

In line with ethnomethodological and interactional-sociolinguistic premises, political discourse is conceived of as 'doing politics', and politicians and journalists present themselves as professional agents who are experts from a factual viewpoint while at the same time also know how to interact with clients. In other words, they do certainty and responsiveness (Fetzer 2008; Simon-Vandenbergen 1996). A linguistic device, which captures the ambivalence of the two basic needs of political discourse in the media, is the contextualization cue *sort of / kind of*, which may make things more fuzzy thus targeting responsiveness in communication, and which may make things less fuzzy thus targeting certainty in communication. And, because of it being a relational concept, it may achieve both: certainty and responsiveness, as is demonstrated with the extracts (28) and (29):

(28) When I left the government I was *not for a moment of the view* that there was a sort of$_1$ <u>return ticket</u> and I haven't changed my mind[…]. I see no point in trying to *if you like* create *another* sort of$_2$ <u>media story</u> it's not gonna happen… [IE 1990a]

(29) Ok. The penny on the income tax will be more than enough to pay for your education, you say, that but it'll only raise about three billion. The sort of <u>spending</u> *you're planning*, with the health service, with *the other bits and pieces*, you're going to need *a vast amount* more [AM 2001f]

Extract (28) displays the less-fuzzy devices "not for a moment of the view", "another sort of media story" and the unambiguous statements "I see no point in trying to create another sort of media story" and "it's not gonna happen" targeting certainty. At the same time, there are the more-fuzzy devices "a sort of return ticket" and "if you like" which target responsiveness by signifying that the

interviewee/interviewer is open for negotiating meaning. Example (28) is of special interest because *sort of* fulfils two different functions within the same contribution.

In (29) there are also more-fuzzy and less-fuzzy devices, viz. "the other bits and pieces" targeting responsiveness by not being too precise and by not threatening the politician's face, and "the sort of spending you're planning" targeting certainty by spelling out clearly what is meant.

Because of their inherent degree of fuzziness, the investigation of *sort of / kind of* in discourse needs to consider their collocations very closely in order to account for their contextualization function. Very often, it is not either a more-fuzzy or a less-fuzzy function but rather a gradable, slightly more-fuzzy or slightly less-fuzzy function.

6. Conclusion

This contribution has examined the theory and practice of *sort of / kind of* in dialogical and monologic political discourse, accounting for collocation, function and distribution. Because of the multiple form-function mapping of *sort of / kind of*, viz. head of NP with postmodifier, adverbial with qualifier function, e.g. hedge, approximator or adapter, the analysis based on their grammatical function within a sentence only does not seem to be appropriate any longer. Instead, *sort of / kind of* are assigned the status of contextualization cue. The definition of contextualization cue is relational and purely functional, accounting explicitly for the cue's embeddedness in context and for co-occurrence.

To count as a contextualization cue, a lexical device must have undergone some development from a narrow-scope modifier of an object or constituent to a wide-scope, metalinguistic modifier and procedural device anchored to the force of a communicative contribution (Aijmer 2002). This generally goes hand-in-hand with a process of grammaticalization and pragmaticalization (Aijmer 1997). Only then can the lexical device be fully indexical, indexing a pragmatic concept, such as plus/minus-fuzziness.

Sort of / kind of have been classified as making a communicative contribution (or one or more of its constitutive parts) more fuzzy or less fuzzy. Depending on local context and co-occurrence, *sort of / kind of* may express either function. If their local context is coloured by devices expressing a higher degree of determinateness, such as definite, demonstrative and interrogative determiners or non-mitigated statements, *sort of / kind of* is more likely to be assigned the status of a less-fuzzy cue. If their local context is coloured by devices expressing a higher degree of indeterminateness, such as expressions of epistemic modality, e.g. probability or possibility, hedges or markers of common ground, *sort of / kind of* is more likely to be assigned the status of a more-fuzzy cue.

In the social context of (spoken) political discourse *sort of / kind of* is less frequent than in informal everyday talk. Here, the cues signify particularized contextual frames, such as particularized discourse topics or negotiation-of-

validity sequences. As less-fuzzy cues they signify certainty, and as more-fuzzy cues they signify responsiveness or attitude.

Notes

The author is grateful to the anonymous reviewers for helpful comments on the first version of this paper. Any remaining errors are exclusively the author's.

1 I would like to thank Peter Bull (University of York, UK) for sharing with me the ten political speeches (2004-2006), the 17 full-length pre-election interviews (1997-2001) and the 2003 interview on Iraq. The four full-length interviews (1990) from the programme *On the Record* (BBC1) were recorded and transcribed by myself.

2 The transcription presented in this paper follows orthographic standards. IR denotes interviewer, IE denotes interviewee, AM denotes audience member taking over the role of IR, and POL denotes the politician delivering a speech. The lexical items over which *sort of / kind of* have scope or which are their postmodifiers are underlined, and expressions relevant to the argumentation are in italics.

3 Gumperz's concept of speech activity is closely connected with Levinson's activity type (Levinson 1979), Linell's communicative project (Linell 1998) and Luckmann's communicative genre (Luckmann 1995).

References

Aijmer, K. (1984), ''Sort of' and 'kind of' in English conversation', *Studia Linguistica*, 38(2): 118-128.

Aijmer, K. (1997), 'I think – an English modal particle', in: T. Swan and O. Jansen (eds.) *Modality in Germanic Languages. Historical and Comparative Perspectives*. Berlin: Mouton de Gruyter. 1-47.

Aijmer, K. (2002), *English Discourse Particles: Evidence from a Corpus.* Amsterdam: John Benjamins.

Andersen, G. (2000), *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language Of Adolescents*. Amsterdam: John Benjamins.

Bar-Hillel, Y. (1998), 'Indexical expressions', in: A. Kasher (ed.) *Pragmatics: Critical Concepts*. London: Routledge. 23-40.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman Grammar of Spoken and Written English.* London: Longman.

Bolinger, D. (1972), *Degree Words*. The Hague: Mouton.

Brown, P. and S.C. Levinson (1987), *Politeness. Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Celle, A. and R. Huart (eds.) (2007), *Connectives as Discourse Landmarks*. Amsterdam: John Benjamins.

Coates, J. (2003), 'The role of epistemic modality in women's talk', in: R. Facchinetti, M. Krug and F. Palmer (eds.) *Modality in Contemporary English*. Berlin: Mouton de Gruyter. 331-348.

Davidse, K., L. Brems and L. De Smedt (2008), 'Type noun uses in the English NP: a case of right to left layering', *International Journal of Corpus Linguistics*, 13(2): 139-168.

De Smedt, L., L. Brems and K. Davidse (2007), 'NP-internal functions and extended uses of the 'type' nouns kind, sort, and type: towards a comprehensive, corpus-based description', in: R. Facchinetti (ed.) *Corpus Linguistics 25 Years On*. Amsterdam: Rodopi. 225-255.

Fetzer, A. (1994), *Negative Interaktionen: Kommunikative Strategien im britischen Englisch und interkulturelle Inferenzen*. Frankfurt: Lang.

Fetzer, A. (2000), 'Negotiating validity claims in political interviews', *Text*, 20(4): 1-46.

Fetzer, A. (2004), *Recontextualizing Context: Grammaticality Meets Appropriateness*. Amsterdam: John Benjamins.

Fetzer, A. (2006), ''Minister, we will see how the public judges you'. Media references in political interviews', *Journal of Pragmatics*, 38: 180-195.

Fetzer, A. (2008), '"And I think that is a very straight forward way of dealing with it". The communicative function of cognitive verbs in political discourse', *Journal of Language and Social Psychology*, 27(4): 384-396.

Greatbatch, D. (1998), 'Conversation analysis: neutralism in British news interviews', in: A. Bell and P. Garrett (eds.) *Approaches to Media Discourse.* Oxford: Blackwell. 163-185.

Greenbaum, S. and R. Quirk (1990), *A Student's Grammar of the English Language*. London: Longman.

Grice, H.P. (1975), 'Logic and conversation', in: P. Cole and J.L. Morgan (eds.) *Syntax and Semantics*. New York: Academic Press. 41-58.

Gumperz, J.J. (1977), 'Sociocultural knowledge in conversational inference', in: M. Saville-Troike (ed.) *Linguistics and Anthropology*. Washington: Georgetown University Press. 191-211.

Gumperz, J.J. (1991), 'Contextualization and understanding', in: A. Duranti and C. Goodwin (eds.) *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press. 229-252.

Gumperz, J.J. (2003), 'Response essay', in: S.L. Eerdmans, C.L. Prevignano and P.J. Thibault (eds.) *Language and Interaction. Discussions with John J. Gumperz*. Amsterdam: John Benjamins. 105-126.

Kay, P. (1984), 'The kind of /sort of construction', *Berkeley Linguistics Society*, 10: 128-137.

Lakoff, G. (1975), 'Hedges: a study in meaning criteria and the logic of fuzzy concepts', in: D. Hockney, W. Harper and B. Freed (eds.) *Contemporary*

Research in Philosophical Logic and Linguistic Semantics. Dodrecht: Fortis. 221-271.

Le Lan, B. (2007), 'Orchestrating conversation: the multifunctionality of *well* and *you know* in the joint construction of a verbal interaction', in: A. Celle and R. Huart (eds.) *Connectives as Discourse Landmarks.* Amsterdam: John Benjamins. 103-116.

Leech, G. (1983), *Principles of Pragmatics*. London: Longman.

Levinson, S.C. (1979), 'Activity types and language', *Linguistics*, 17: 365-399.

Levinson, S.C. (1983), *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S.C. (2000), *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge: MIT Press.

Linell, P. (1998), *Approaching Dialogue*. Amsterdam: John Benjamins.

Longman Dictionary of Contemporary English (2003). Longman: Harlow.

Luckmann, T. (1995), 'Interaction planning and intersubjective adjustment of perspectives by communicative genres', in: E. Goody (ed.) *Social Intelligence and Interaction: Expressions and Implications of the Social Bias in Human Intelligence.* Cambridge: Cambridge University Press. 175-188.

Mauranen, A. (2004), 'There's a little bit different... Observations on hedges in academic talk', in: K. Aijmer and A.B. Stenström (eds.) *Discourse Patterns in Spoken and Written Corpora.* Amsterdam: John Benjamins. 173-197.

Prevignano, C. and A. di Luzio (2003), 'A discussion with John J. Gumperz', in: S.L. Eerdmans, C.L. Prevignano and P.J. Thibault (eds.) *Language and Interaction. Discussions with John J. Gumperz*. Amsterdam: John Benjamins. 7-29.

Prince, E., J. Frader and C. Bosk (1982), 'On hedging in physician-physician discourse', in: R.J. Di Pietro (ed.) *Linguistics and the Profession*. Norwood: Ablex. 83-97.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Grammar of Contemporary English.* London: Longman.

Schiffrin, D. (1987), *Discourse Markers*. Cambridge: Cambridge University Press.

Searle, J.R. (1969), *Speech Acts*. Cambridge: Cambridge University Press.

Simon-Vandenbergen, A.M. (1996), 'Image-building through modality: the case of political interviews', *Discourse and Society*, 7: 389-415.

Traugott, E.C. (1995), 'Subjectification in grammaticalization', in: D. Stein and S. Wright (eds.) *Subjectivity and Subjectivisation*. Amsterdam: John Benjamins. 31-54.

Ungerer, F. (1988), *Syntax der englischen Adverbialien.* Tübingen: Niemeyer.

Verhagen, A. (2005), *Constructions of Intersubjectivity*. Oxford: Oxford University Press.

Appendix

Table 1a: The political interviews

Date	Interviewee + Interviewer
2003 Iraq	T. Blair (Labour PM) + J. Paxman
2001a PEI	T. Blair (Labour PM) + D. Dimbleby
2001b PEI	T. Blair (Labour PM) + J. Dimbleby
2001c PEI	W. Hague (Conservative Opposition) + D. Dimbleby
2001d PEI	W. Hague (Conservative Opposition) + J. Dimbleby
2001e PEI	Ch. Kennedy (Liberal Democrats) + D. Dimbleby
2001f PEI	Ch. Kennedy (Liberal Democrats) + J. Dimbleby
1997a PEI	J. Major (Conservative PM) + M. Buerk
1997b PEI	J. Major (Conservative PM) + J. Dimbleby
1997c PEI	J. Major (Conservative PM) + J. Paxman
1997d PEI	T. Blair (Labour Opposition) + D. Dimbleby
1997e PEI	T. Blair (Labour Opposition) + J. Dimbleby
1997f PEI	T. Blair (Labour Opposition) + D. Frost
1997g PEI	T. Blair (Labour Opposition) + J. Paxman
1997h PEI	P. Ashdown (Liberal Democrats) + D. Dimbleby
1997i PEI	P. Ashdown (Liberal Democrats) + J. Dimbleby
1997k PEI	P. Ashdown (Liberal Democrats) + D. Frost
1997l PEI	P. Ashdown (Liberal Democrats) + J. Paxman
1990a OTR	M. Heseltine (Conservative) + J. Dimbleby
1990b OTR	J. Prescott (Labour Opposition) + J. Dimbleby
1990c OTR	J. Cunningham (Labour Opposition) + J. Dimbleby
1990d OTR	J. Gummer (Conservative Minister) + J. Dimbleby

Table 2a: The political speeches

Date	Politician
2006a	G. Brown (Labour Party Conference)
2006b	D. Cameron (Conservative Party Conference)
2006c	M. Campbell (Liberal Party Conference)
2005a	D. Cameron (Conservative Party Conference)
2005b	D. Davis (Conservative Party Conference)
2005c	L. Fox (Conservative Party Conference)
2005d	Ch. Kennedy (Liberal Democrat Party Conference)
2005e	M. Rifkind (Conservative Party Conference)
2005f	T. Blair (Labour Party Conference)
2004	M. Howard (Conservative Party Conference)

"So er I just sort I dunno I think it's just because…": A corpus study of *I don't know* and *dunno* in learners' spoken English

Karin Aijmer

University of Gothenburg

Abstract

The aim of this study is to investigate if I don't know *and* (I) dunno *are used in the same way by learners and by native speakers. The data come from the Swedish component of the recently compiled LINDSEI corpus of spoken learner texts. A comparison is made with native speaker data from the LOCNEC corpus.* I don't know (dunno) *is characteristically multifunctional and needs to be described with regard to parameters like involvement (epistemic and affective stance), speech-management and politeness.* I don't know (dunno) *can also be used to take or yield the turn or to mark the opening or closing of a topic. The comparison between native speakers and learners shows that learners overwhelmingly use* I don't know (dunno) *as a speech management signal. Native speakers use* I don't know *mainly to avoid asking questions in a direct way.*

1. Introduction

The following extract illustrates a pragmatic marker which is informal or non-standard:

(1) <B> so I was helping out there and then they gave me free food <laughs> and I just I *dunno* I just stayed and lived with them <\B>
<A> what's a . Thai: breakfast look like <\A>
<B> they: eat: sticky rice <\B>
<A> for breakfast <\A>
<B> yeah . with I *dunno* I can't remember what it's called some m= m= mashy thing with l= small fish <\B>
<A> mhm <\A>
<B> and . *dunno* cold<?> . with the sticky rice <\B> (LINDSEI)

(I) dunno is a reduced form of *I don't know*. *I don't know (I dunno)* seems to be semantically transparent. However, *I don't know* (*I dunno)* is not always used in its literal meaning. Recent research has drawn attention to the fact that *I don't know* can also be used when the speaker is actually able to supply the information (Tsui 1991; Scheibman 2000; Kärkkäinen 2003; Diani 2004; Grant forthcoming).

I don't know can express different attitudes and has much in common with modal particles in other languages. In the extract above *(I) dunno* seems to mean uncertainty. The pragmatic function to express uncertainty can easily be understood from the literal meaning of *I don't know*. According to Beach and Metzger (1997: 569), "marking uncertainty and doubt appears to be a central feature of a variety of claims of insufficient knowledge". The dictionary notes, for instance, that *I don't know* is used 'for saying that you do not completely agree'. "It'll be boring. Oh I don't know. It might be fun" (*Macmillan Dictionary of Advanced Learners*). Östman was one of the first linguists to discuss the function of this construction (although only briefly):

> There is one other pragmatic construction involving the segment *know* which I would like briefly to mention in this connection, since it can also have a floor-yielding function. This is the use of *I don't know* to express uncertainty (as opposed to 'I have no knowledge'), paraphrasable as "I don't really know what to think about that", or, "I don't really have anything more to say about this", or some similar qualification. (Östman 1981: 27)

Östman discusses mainly the floor-yielding function of *I don't know* which coincides with its use at the end of a turn. *I don't know* can also be a prefatory frame functioning as a stepping-stone to what follows in the conversation.

The speaker in extract (1) above is a learner of English who uses *(I) dunno* when he or she has difficulty finding the right words. It illustrates the type of data which I use in this paper. Learners have been shown not to use all the functions of pragmatic markers or to use them with different frequencies from native speakers (Müller 2005). This makes it interesting to study if *I don't know* (and *I dunno*) are used in the same way by learners and by native speakers. Previous work, for example by Tsui (1991) and Diani (2004), has shown that *I don't know* is, above all, important for cooperation and maintaining the social balance in conversation. However, it does not follow that it is used by learners for the same ends.

The outline of this paper is as follows. In section 2, I discuss the material, followed, in section 3, by a description of the distribution of *I don't know* in the data. The positional variability of *I don't know* is discussed in section 4. Section 5 provides the background for dealing with the multifunctionality of *I don't know* (and other pragmatic markers). *I don't know* has a large number of different functions as a pragmatic marker which will be analysed in sections 6-8. The final section (section 9) summarises the discussion and draws some conclusions about how advanced learners use *I don't know* and *(I) dunno*.

2. The material

Learners are a heterogeneous category of language users. It is therefore difficult to speak about learners in general. There are e.g. Swedish learners of English,

German learners of English, etc. who can be expected to use markers differently depending on their native language. My study will compare Swedish advanced learners of English with native speakers.

The data come from the Swedish component of the recently compiled LINDSEI corpus of spoken learner texts, the sister project of the SWICLE corpus of written English (Aijmer 2004). A comparison is made with data from the LOCNEC corpus compiled by De Cock – a mirror image of the learner corpus but with young native speakers of English as interviewees instead of learners (De Cock 2004). The LOCNEC corpus contains 117,147 words (De Cock 2004) compared with 144,355 words in the Swedish component of LINDSEI. The informants were interviewed and encouraged to speak about a topic such as travels, plans for the future, or university life. In addition, the interviewees were asked to tell a short story based on a picture.

3. The distribution of *I don't know*

I don't know is one of the most frequent personalised epistemic pragmatic markers, although it is not as frequent as *I think.* It occurs with a normalised frequency of 1,163 tokens per million words in English conversation compared with only 51 tokens in writing (Rühlemann 2007: 117). There are 138 examples of *I don't know* in LINDSEI (including *I don't know a word*, *I don't know about* and *I don't know* with an interrogative word) and 96 examples of the reduced form *dunno.* This is compared with 210 examples in LOCNEC (in LOCNEC the difference between *I don't know* and *I dunno* has not been noted in the transcriptions).

I don't know is, above all, frequent in the negative form. The corresponding positive form *I know* occurs 146 times in LOCNEC and 42 times in LINDSEI. In the *British National Corpus* (BNC) there were 6,870 examples of *I don't know* compared with 8,785 tokens of *I know.* A comparison can also be made with *I think* and *I don't think.* There were 8,912 examples of *I think* compared with only 1,938 examples of *I don't think* in the BNC.

The variant *I dunno* occurs in different surface forms such as *dunnu* or *dunnow* (both discussed on the internet www.urbandictionary.com/define.php?term=i+don't+know; www.urbandictionary.com/define.php?term=dun no). *I dunno* is the result of semantic bleaching and can be regarded as more 'pragmaticalised' than *I don't know.*[1] The reduced form is used in both British and American English in variation with *I don't know* (cf. Scheibman 2000 for American English).

The high frequency of *I don't know* in learner texts is confirmed in related work on spoken learner English. For example, *I don't know* occupied the first rank among three-word sequences in the study by De Cock of French learners' use of recurrent sequences (De Cock 2004: 232) where it is more frequent than *and it was* and *and er well.*

I don't know is not semantically and syntactically compositional but a "preferred sequence of words" (De Cock 2004) or a "lexical bundle" (Biber et al.

1999). For example, Biber et al. (1999: 1002) discuss *I don't know* and combinations with *I don't know* as lexical bundles or "recurrent discourse building blocks", i.e. bundles of words that show a statistical tendency to co-occur (1999: 991).

A distinction needs to be made between *I don't know* and other lexical bundles with *I don't know*, in particular the literal use illustrated in example (2).

(2) <A> so in what way do you think your flat reflects your personality <\A>
<B> <begin laugh> *I don't know* I still live with my parents (LINDSEI)

I don't know expresses the speaker's reluctance or unwillingness to answer the question directly.

(3) <A> what kind of music <\A>
<B> erm .. *I don't know how* to characterise it's: sort of pop music (LINDSEI)

A large number of my examples of *I don't know* are of the type illustrated by (3). The following four-word bundles were found in the two corpora (table 1).

Table 1: *I don't know* followed by an interrogative word in LOCNEC and LINDSEI

	LINDSEI			LOCNEC
Bundle	*I don't know*	*(I) dunno*	*I don't know* and *(I) dunno*	
I don't know what	5	5	10	9
I don't know why	4	1	5	7
I don't know if	14	10	24	22
I don't know how	5	3	8	8
I don't know whether	3	-	3	-
I don't know+ other wh-words (when, where, who, which)	3	-	3	4
Total	34	19	56	50

I don't know if was by far the most frequent of the bundles, but other combinations were frequent as well.

The 'bound tokens' of *I don't know* (*dunno)* illustrated in table 1 (cf. Pichler unpublished) have not been analysed since they are not pragmatic markers. Other examples which have not been included because *I don't know* takes its literal meaning include *I don't know* with a referential object (*I don't know the word*, *I don't know him*, *I don't know a lot about it*) and with adverbials in some cases (*I don't know yet*, *I don't really know*). This leaves us with 79 examples of *I don't know* and 77 of *(I) dunno* as a pragmatic marker compared with 141 examples from LOCNEC.

There is considerable individual variation when we look at the use of *I don't know* and *(I) dunno* (see appendix for details). Some learners (nine learners out of 50) do not use *I don't know* (or *I dunno*) at all. Five learners stood out because they only used *I dunno* and 17 learners only used *I don't know*. Two speakers were in fact responsible for most of the examples of *dunno*. They were also the speakers who used the pragmatic marker most frequently (twelve and thirteen times, respectively, in raw frequencies).

4. *I don't know* and position

Assigning a function to *I don't know* needs to be done on the basis of the linguistic analysis of the examples in the corpus (e.g. the relationship between the function and different positions in the turn). *I don't know* as a pragmatic marker is usually not integrated in the grammatical structure and is found in many different positions. Table 2 shows the distribution of *I don't know* and *(I) dunno* in the two corpora.

Table 2: The different positions of *I don't know (dunno)* in LINDSEI and LOCNEC (percentages refer to relative frequencies)

	LINDSEI			LOCNEC
	I don't know	*(I) dunno*	*I don't know* and *(I) dunno*	
Initial	19 (24.1%)	17 (22.1%)	**36 (23.1%)	**60 (42.6%)
Mid	25 (31.6%)	35 (45.5%)	60 (38.5%)	37 (26.2%)
End	31 (39.2%)	24 (31.2%)	55 (35.3%)	34 (24.1%)
Alone	4 (5.1%)	1 (1.3%)	5 (3.2%)	10 (7.1%)
Total	79 (100%)	77 (100.1%)	156 (100.1%)	141 (100%)

** indicates that the difference is highly significant ($p<0.001$)

I don't know occurs initially in the turn (also after a conjunction). When *I don't know* is placed in medial position (any position which is neither initial nor final) it can modify a single constituent rather than the whole clause (see example (4)).[2]

(4) <B> mm .. it's not that small but it's smaller and eh . *I don't know* .. more . cold atmosphere there I think so . I love Gothenburg <laughs> <\B> (LINDSEI)

The turn-initial position is significantly more frequent in LOCNEC than in LINDSEI. Moreover, *I don't know* was found more often at the end of the

utterance in LINDSEI than in LOCNEC (at a TRP or transition relevance place according to Conversation Analysis) (Sacks et al. 1978: 12). *I don't know* can also be used alone. *I dunno* was frequent in mid position also in comparison with *I don't know* in LINDSEI and in LOCNEC.

The position of *I don't know* in the clause is pragmatically or interactively motivated rather than syntactically and is therefore an important cue to its function (Corum 1975). Depending on its position in the turn it can, for instance, be used to take the turn, to keep the turn, or to yield the floor to another speaker.

5. *I don't know* and multifunctionality

The article by Tsui (1991) on the pragmatic functions of *I don't know* has sparked off a discussion of how *I don't know* is used. Tsui noted that *I don't know (dunno)* is a common reply when one does not want to give an answer even if one has the information. One reason may be that the speaker does not want to disagree with the hearer. However, *I don't know* does not have a single function but is characterised by its broad spectrum of uses. Because of this, it is important to discuss how the use of *I don't know* is constrained by the context. The multifunctionality of *I don't know* should be understood in the sense suggested by Östman: "the concept 'multifunctionality' receives quite a new meaning when it is seen in the light of the different functions a pragmatic particle has in relation to a systematic set-up of independently defined functions and contexts" (Östman 1995: 106-107). Östman has described not only textual and interactional factors which are important to analyse pragmatic markers but also parameters such as cultural coherence, involvement and politeness which are pragmatic in a wider sense.

(Cultural) coherence refers to the cultural and social norms constraining what we say. This parameter explains, for example, that pragmatic markers have discourse marking functions to create discourse coherence or, in Östman's words, they provide the "grease" that makes the conversation flow more smoothly.

I don't know can also have functions relating to 'involvement'. Involvement "is concerned with how to express or not express feelings, attitudes and prejudices" (Östman 1995: 104). *I don't know* and its close 'relatives' *I think*, *I believe* express epistemic or affective stance to the hearer or to what is said.

The third parameter is politeness, defined as 'the interactional constraints we follow when establishing, maintaining and breaking interpersonal relations' (Östman 1995). Since Tsui's (1991) pioneering article on *I don't know*, the marker has been associated with deference and politeness. In Tsui's data, *I don't know* expresses 'polite avoidance' and can be used to ward off the negative consequences of face-threatening acts such as requests or compliments. However, this function was less apparent in my data which may be due to the fact that the interviewees are learners but also to the special constraints imposed by the interview situation: in interviews *I don't know* is frequently found in answers to questions with the function of signalling unwillingness to answer the question.

The conversation, or the interview in this case, takes place in real time. The temporal constraint can be assumed to cause special problems for learners who have to struggle more to get their messages across. An important function of many pragmatic markers is that they can be used to gain time for planning and self-correction. The 'scarcity of time' constraining the production of speech in real time (Rühlemann 2007: 49) is also reflected in the use of lexical bundles rather than a single element. *I don't know* for example is both used alone as a pragmatic marker and embedded in a lexical bundle with *and*, *but*, *I think* etc. to show how different units of the text are connected. This function is closely related to the speech management function (cf. Östman's coherence function).

When a pragmatic marker is multifunctional it potentially has meaning with regard to a number of (textual and interactive) functions and contexts including the parameters discussed by Östman. A particular function can, however, be highlighted in the discourse. *I don't know* can, for instance, have the polite function of conveying deference or of softening an assertion or it can have mainly a speech management function to facilitate production and processing by adding to the coherence of the discourse. It can signal uncertainty, reluctance to commit oneself, or discomfort (involvement). However, there is much ambiguity and variability and "the ultimate 'meaning' [of the communicative behaviour of the pragmatic particles] will also be influenced by the contextual and interactional effect the message in question has" (Östman 1995: 105). For example, the meaning of *I don't know* is influenced by its role to take or yield the turn, the speaker's attitude to the hearer or to the text, whether the speaker wants to continue the topic or not, and whether *I don't know* is embedded in a lexical bundle.

The speech management function is particularly important when we discuss *I don't know.* As indicated in extract (1), *I don't know* helps the speaker to manage the processing of the utterance on-line by providing bonus time. In section 9, I will return to a discussion of how the functions of *I don't know* are constrained by the fact that we are referring to learners.

6. *I don't know* as a floor-yielding or topic-closing signal – 'the speaker has nothing more to say'

When *I don't know* is placed at the end of the turn or utterance it has the function of yielding the floor or fulfilling the desire of the interviewee 'to close a topic' (Pichler unpublished) in addition to its attitudinal function to express uncertainty or lack of responsibility ('I have nothing more to say – the floor is yours if you want it'). *I don't know* was often used with *maybe (perhaps*, *I think*, *might*, etc.) in the same sentence, further emphasising the tentativeness of the speaker's contribution:

(5) <B> and she told me that my accent was quite <begin laugh> messed up <end laugh> and I think maybe it is *I don't know* <\B> (LINDSEI)

It is not always the case in interviews that a speaker change occurs after *I don't know*, the current speaker may choose to continue:

(6) <B>yeah *I think that might* have something to do with it erm *I don't know* I've just always felt more comfortable in Ireland and that's maybe where I. I fit in and <\B> (LOCNEC)

I don't know in the potential topic-closing function was typically used by both learners and native speakers. In this function it was followed by a new turn in 14 examples out of 31 in LINDSEI. The corresponding figure for *dunno* was ten out of 24 examples. In LOCNEC, *I don't know* in the topic closing function was followed by a new turn in half of the examples (17 out of 34 examples). The topic-closing function was the most frequent function associated with *I don't know* in the learner corpus (cf. section 8, table 4).

The topic-closing or floor-yielding signal is only possible if *I don't know* points backwards and is placed at the end of the turn (or utterance). Kärkkäinen (2003) only found a few examples with *I don't know* in final position with interactional function in the *Corpus of American English* which provided her data (five conversational samples from the *Santa Barbara Corpus of Spoken American English*): "Even *I don't know*, which is one of the most mobile epistemic phrases and as a consequence has acquired more versatile interactional functions, occurs in turn-final position rather infrequently" (Kärkkäinen 2003: 94). It is therefore possible that this function is especially frequent in interviews.

7. *I don't know* as a preface – 'there is more to come'

I don't know functions, above all, as a preface which, in conversation-analytic terms, is 'designed to project that there is more to come'. Schegloff (1996) refers to this use as a "prefatory epistemic disclaimer". As a preface in thematic position, *I don't know* occurs in an interactionally and pragmatically interesting position where it marks a stepping-stone to what comes next, signalling the starting-point of a speaker perspective. For example, it can mark a new topic (or topic shift) or be used to take a turn.

The subjective or stance meaning is particularly important (cf. Östman 1995 'involvement'). Biber et al. draw attention to the relationship between the speaker and the proposition (epistemic stance):

> In most cases, speakers and authors first identify their personal perspective – their attitude towards the proposition, the perspective that it is true from, or the extent to which the information is reliable – thereby encouraging listeners and readers to process the following propositional information from the same perspective. (Biber et al. 1999: 971)

I don't know can, for instance, signal the speaker's uncertainty, an apologetic or defensive attitude to the message or to the hearer, as well as reluctance or discomfort if the message is embarrassing. In (7), the speaker's attitude is, for example, defensive or apologetic ("I haven't really thought about it"):

(7) <A> that's right would you like to have your own portrait painted . or would it be a little bit eh . you know <\A>
<B> *I don't know* really I haven't really thought about it it depends on who is painting it [I think <\B> (LINDSEI)

I don't know as a preface signals hesitation which I regard as a structural or speech management function since the pausing occurs between units in the discourse and contributes to coherence. The function of *I don't know* can be explained in temporal terms: the speaker makes a pause in the discourse to think about what to say next. However, *I don't know* as a preface can have several functions which can be present simultaneously. For example, Pichler (unpublished: 2) analyses some uses of *I don't know* as having a 'subjective-textual' function, it "can initiate a turn and at the same time hedge its content". An example would be:

(8) <A> .. and em .. will you have lots of things in your flat or keep it . you know will you keep it very simple <\A>
<B> mm: <begin laugh> *I don't know* at first it will probably look like it really <giggles> occupied <end laugh> <giggles> <\B> (LINDSEI)

The speech management function is particularly important when we discuss *I don't know* as a preface since so much planning takes place at the beginning of the utterance or the turn. In interaction-based studies of language, constituents in the thematic position are regarded as 'emerging-in-time as the speaker begins to construe his or her turn or turn-component' (Auer 1996: 307). As a result, *I don't know* can be used as a structural signal filling a gap needed for processing purposes.

(9) <B> eh sorry a what [<laughs> <\B>
<A> [a macho you know a very sort of male <\A>
<B> er <\B>
<A> erm <\A>
<B> <tuts> *I don't know* well in this film . that part of the culture was emphasised <\B> (LINDSEI)

In section 7.1, I will discuss *I don't know* as a mitigator or a hedge. In section 7.2, I will discuss functions which can be realised by lexical bundles.

7.1 *I don't know* hedging opinions

As pointed out by other scholars, *I don't know* is used for politeness or face-saving functions (Östman 1995; Tsui 1991; Diani 2004; Grant forthcoming). For example, just like other modal particles (e.g. *I think*), *I don't know* can be used as a hedge to mark uncertainty or tentativeness. *I don't know* typically co-occurs with other epistemic markers such as *maybe*, *just* or *really:*

(10) <B> hm *I don't know* . maybe <begin laugh> I can go back to Bulgaria I [<end laugh> <\B> (LINDSEI)

According to Brown and Levinson (1987: 116), the speaker "may choose to be vague about his own opinions, so as not to be seen to disagree". The speaker avoids expressing an opinion directly by using the 'safely vague' *I don't know* thus avoiding disagreement and protecting his or her positive face needs.

(11) <B> that's like really nice there and <XX> seem to see you know <X> walk up the spine and .. *I don't know* it's nice it's modern really <\B> (LOCNEC)
(12) *and I don't know* it w= it w= it was just amazing it was just sort of erm .. <X> it it was really sort of . grim and grotesque but but because you never saw anything happen <\B> (LINDSEI)
(13) <B> *don't know .. erm I don't know* it was really strange cos I put I mean I s= filled out all the[i:] er UCAS forms and decided where to go <\B> <A> mm <\A> (LOCNEC)

Native speakers used *I don't know* to hedge their utterances more often than the learners (see section 8, table 4). However, above all, native speakers used *I don't know* in lexical bundles more frequently than learners.

7.2 *I don't know* in lexical bundles

I don't know often combines with other elements especially in initial position (cf. Diani 2004). *Well I don't know*, *oh I don't know*, *and I don't know*, *but I don't know*, *so I don't know*, *I don't know I think*, *I don't know I mean* are preferred sequences or lexical bundles in the terminology I have used above. Just like 'single markers' they help the speaker to process, plan and execute utterances. In table 3 some frequent bundles with *I don't know* are compared.

In a larger corpus such as the BNC, *well I don't know*, *I don't know I mean*, *oh I don't know* and *I don't know you know* are extremely frequent, indicating that they have been routinised and are important in discourse (cf. Diani 2004 who analysed the functions of *I don't know* with discourse markers in the *Cobuild/Birmingham Spoken Corpus*).[3]

Table 3: Bundles of pragmatic markers with *I don't know* in LINDSEI and in LOCNEC

	LINDSEI			LOCNEC
Bundle	*I don't know*	*(I) dunno*	*I don't know* and *(I) dunno*	
and I don't know	3	1	4	4
but I don't know	2	3	5	8
so I don't know	4	2	6	2
(oh) well I don't know	3	1	4	6
I don't know I think	-	4	4	10
I don't know I mean	-	1	1	6
I don't know I don't think	1	-	1	2
I don't know you know	-	-	-	1
oh I don't know	1	2	3	1

Both learners and native speakers use the bundles in table 3 (and other less frequent ones). For example, native speakers used *I don't know* before *I think*, *I mean* and *I suppose* and after *but* or *well* with new meanings.

In my data, *but I don't know* and *well I don't know* had different meanings. In the following example *but I don't know* is both a hedge and a preface (topic-changing or signalling a new turn).

(14) <B> erm I was thinking about doing an M A <\B>
<A> an M A . here <\A>
<B> yeah probably <\B>
<A> in language studies <\A>
<B> yeah <\B>
<A> mhm <\A>
<B> *but I don't know* it depends <X> getting my degree <\B> (LOCNEC)

Well I don't know, on the other hand, is used to signal polite disagreement:

(15) <B> [yeah .. yeah I don't think I'll be going again next summer <X> it's turned out very expensive but it was [worth it <\B>
<A> [<X> is it <\A>
<B> *well I don't know* our flight was four hundred pounds <\B>
<A> mhm <\A>
<B> which was a lot and then being there for two weeks and it wasn't normally I go on like say a beach holiday <\B> (LOCNEC)

Well I don't know suggests the speaker's unwillingness to admit that the flight which cost four hundred pounds was really quite expensive.

Similarly, *oh (I) don't know* can preface disagreement as in example (16). (See Diani 2004: 166: "(*Oh) I don't know* has the effect of avoiding a forthcoming disagreement".)

(16) <B> *oh don't know* I don't think so it my impression is that <sigh> everything is focused around work erm work is very important and you can almost never take a day off: erm go anywhere because you would lose opportunities to: to get a better situation mm . work-wise .. [<XX> <\B> (LOCNEC)

Bundles such as *I don't know I mean* or *I don't know I think* are used when the speaker suddenly recollects something in the middle of the discourse or wants to clarify, explain or correct something:

(17) <B> to be I mean <X> just in the setting I mean in <title of a play> by setting it in Soho it made it that much more relevant to us <\B>
<A> [oh yes <\A>
<B> [than if it was set in *I don't know . I think* it was in Berlin or somewhere <\B> (LOCNEC)

Lexical bundles help the speaker to process the utterance since they come 'ready-made'. Learners use them less frequently and not in the same way. *So I don't know* is, for instance, used only by learners as a hedge and to mark topic-closure (yielding the turn):

(18) <B> erm <breathes> I don't know. erm .. well I haven't actually thought of it [<laughs> <\B>
<A> [<laughs> <\A>
<B> *so I don't know*. erm <\B> (LINDSEI)

8. *I don't know* as a speech management or coherence signal

I don't know has discourse coherence functions in the sense of Östman (1995). The term 'speech management signal' which I have used is borrowed from Allwood et al. (1990: 3) who define the concept 'speech management phenomenon' (SM) as follows: "The concept of SM involves linguistic and other behavior which gives evidence of an individual managing his or her own communication while taking his or her interlocutor into account".

The authors make a distinction between choice-related and change-related speech management functions. *I don't know* can be analysed from both perspectives. The choice-related function "is to enable the speaker to gain time for processes having to do with the continuing choice of content and types of structured expression. In particular, such processes can be connected with

prompting of memory, search of memory, hesitation and planning" (Allwood et al. 1990: 10-11).

I don't know can be used with the speech management function to gain time for planning ahead, also when it occurs first in the turn. The speech management function is further signalled by repetition, co-occurrence with pauses, and clustering of pragmatic markers:

(19) <A> mm what about in your: free time . if you have much . what <\A>
<B> er *well I suppose I dunno* I've just started playing football again . so eh I . did play a lot when I was younger <\B> (LINDSEI)

The speaker has to answer the question (what do you do in your free time?) but needs time to make up his or her mind about what to say. *I don't know* (*I dunno)* rather than some other marker expresses the speaker's involvement in the discourse (neither agreement nor disagreement). It can, for instance, reflect the fact that the speaker is 'beating around the bush', especially if the answer is embarrassing to the speaker and the hearer. *I don't know* points neither forwards nor backwards in the discourse but has the structural function of filling a gap in the discourse, signalling that the speaker does not know what to say yet.

In (20), *dunno* fills a slot in the discourse before the speaker comes up with the name of the country he or she has in mind:

(20) <A> so you don't get these monsoon periods <\A>
<B> you do get the monsoon but not as . much or as hard as in . *dunno* India [for <\B> (LINDSEI)

The speech management or coherence function can also be expressed by a lexical bundle expressing tentativeness (*I don't know I don't think*):

(21) yeah.. *I don't I don't know I don't think* I particularly preferred either (LOCNEC)

The change-related function, on the other hand, has to do with self-repair or with self-interruption. *I don't know* has this function when it does not modify the whole sentence but has narrow scope and modifies a constituent (a single word or a phrase). In (22), for example, the speaker changes the direction of his or her talk in the middle of the turn and restarts. *I don't know* rather than *well* or some other speech management signal conveys that the speaker reluctantly and after inner consultation chooses how to continue.

(22) <B> er and s= it seems to me it mocks . the[i:] the the . <tuts> *I dunno* it mm does something <begin laughter> like that <end laughter> <\B> (LINDSEI)

In the following example, the speaker stops in the middle of the utterance and uses *I don't know* to mark a restart also signalled by 'deleting' *when* and substituting it with *because*. *I don't know* signals not only a change in direction of the talk but the speaker's unwillingness to commit him- or herself.

(23) no when *I don't know* because you're with people your own age (LOCNEC)

The speech management function was overwhelmingly associated with the learners (see table 4). Learners used this function almost three times as often as native speakers. *Dunno* had the same frequency as *I don't know* in this function.

By way of summing up, I show the functions of *I don't know* and *dunno* in the two corpora (table 4). The semantic classification is only coarse-grained. It does not, for example, show the difference between marking uncertainty and avoiding commitment (both have to do with involvement and stance and have been classified as hedging). Notice that approximators (the use of hedges before numerals) are 'hidden' in the category of hedges.

As I have emphasised above, *I don't know* can have meaning with regard to several parameters (involvement, coherence or speech management, politeness). Other parameters are turn-taking (taking and yielding the turn) and topic structure (opening and closing a topic). Moreover, *I don't know* can have several meanings simultaneously. *I don't know* can, for instance, occur first in the turn with hedging and speech-management functions:

(24) <B> <tuts> erm no cos I think he likes abstract art as [well <\B>
<A> [uhu [mhm <\A>
<B> [erm but erm .
eh *I don't know* erm . I haven't thought of it actually <\B> (LINDSEI)

As shown in table 4, there are clear differences between learners and native speakers. *I don't know* was, for instance, more frequent with the floor-yielding or topic-closing function in the LINDSEI corpus (although this is a common function in both groups of speakers). The most striking result of the comparison is, however, the overwhelming use of the speech-management function in the learner corpus. On the other hand, native speakers used *I don't know* significantly more frequently than learners with the function of avoiding a direct answer to a question.

Learners used both *I don't know* and *(I) dunno*. However, they did not use them in the same way. (*I*) *dunno* was, for instance, more frequent than *I don't know* in the function hedge + speech management. On the other hand, *I don't know* was used more often to avoid answering a question directly.

Table 4: The frequencies of different functions of *I don't know* and *(I) dunno* in LINDSEI and LOCNEC (relative frequencies are indicated in percentages)

	LINDSEI			LOCNEC
	I don't know	*(I) dunno*	*I don't know and (I) dunno*	
Topic closing /hedging	31 (39.2%)	24 (31.2%)	55 (35.3%)	32 (22.7%)
Speech management (coherence)	15 (19%)	15 (19.5%)	**30 (19.2%)	**7 (5%)
Hedge (involvement)	6 (7.6%)	7 (9.1%)	13 (8.3%)	14 (10%)
Hedge +speech management	9 (11.4%)	24 (31.2%)	33 (21.1%)	24 (17.1%)
Inability to answer / avoiding a straight answer (politeness)	15 (19%)	5 (6.5%)	***20 (12.8%)	***57 (40.4%)
Prefacing disagreement (politeness)	3 (3.8%)	2 (2.6%)	5 (3.2%)	7 (5%)
Total	79 (100%)	77 (100.1%)	156 (99.9%)	141 (100.2%)

** indicates that the difference is highly significant ($p<0.001$) *** indicates that the difference is significant at the $p<0.0001$ level

9. Conclusion

The results of the present study are relevant for the study of pragmatic markers in general and, in particular, for pragmatic markers which can be characterised as modal particles, to use a term which has been used to describe similar phenomena in other languages.

It has been argued that pragmatic markers are characteristically multifunctional and need to be described with regard to both the immediate situational context and the embeddedness of speech in a social and cultural context. Besides the textual and interactive function, there are certain broad dimensions to which pragmatic markers must be oriented, such as coherence, involvement and politeness, and which therefore play an important role in how pragmatic markers are described. Involvement has to do with speaker and hearer (epistemic and affective) stance. *I don't know* is, for example, oriented to the speaker and has a stance-marking function (epistemic uncertainty). The speaker's motivation for expressing uncertainty can be to hedge or soften an opinion or assertion. Cultural/discourse coherence refers to social conventions and norms and is

associated with the discourse marking function of providing the "grease" which makes the conversation run smoothly. Pragmatic markers can, for instance, also mark the end of a topic (or the start of a new one) although these functions usually co-occur with other functions.

The focus of this article has been on advanced learners and how they use *I don't know.* When we study learners we may get a different picture of what a pragmatic marker 'means' from when we study native speakers. Both the number of functions and the type of function depend on who the speakers are and the type of speech situation (for instance, if it is an interview or a conversation).

The results of the investigation have shown that learners overwhelmingly use *I don't know* as a speech management signal rather than for other functions. The scarcity of time in spoken communication is a considerable problem for learners since planning starts from scratch and they cannot be expected to have internalised and routinised structures to the same extent as native speakers.

Notes

1 The most pragmaticalised form, *dunno* without a personal pronoun, was found in only eight examples. *Don't know* without a subject occurred three times.

2 Such examples would be analysed by Kärkkäinen (2003: 65) in prosodic terms as occurring intonation unit-initially rather than grammatically: "when an utterance is not qualified right at the outset, the qualification is 'precision-timed' to come immediately before the relevant issue, simultaneously starting a new intonation unit at that point".

3 In the *Cobuild/Birmingham Spoken Corpus* (about 2 million words) *well I don't know* was the most frequent combination (198 examples). *Oh I don't know* occurred less often (38 examples), for *I don't know I mean* (or *I mean I don't think*) there were 85 examples and *I don't know you know* 23 examples.

References

Aijmer, K. (2004), 'Pragmatic markers in spoken interlanguage', *Worlds of Words. A Tribute to Arne Zettersten. Nordic Journal of English Studies Special Issue*, 3(1): 173-190.

Allwood, J., J. Nivre and E. Ahlsén (1990), 'Speech management: on the non-written life of speech', *Nordic Journal of Linguistics*, 13: 3-48.

Auer, P. (1996), 'The pre-front field in spoken German and its relevance as a grammaticalization position', *Pragmatics*, 6(3): 295-322.

Beach, W.A. and T.R. Metzger (1997), 'Claiming insufficient knowledge', *Human Communication Research*, 23(4): 562-588.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *The Longman Grammar of Spoken and Written English.* London: Longman.

Brown, P. and S.C. Levinson (1987), *Politeness. Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Corum, C. (1975), 'A pragmatic analysis of parenthetic adjuncts', in: R.E. Grossman, L.J. San and T.J. Vance (eds.) *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society.* Chicago: Chicago Linguistic Society. 131-141.

De Cock S. (2004), 'Preferred sequences of words in NS and NNS speech', *Belgian Journal of English Language and Literatures (BELL),* New Series 2: 225-246.

Diani, G. (2004), 'The discourse functions of *I don't know* in English conversation', in: K. Aijmer and A.-B. Stenström (eds.) *Discourse Patterns in Spoken and Written Corpora*. Amsterdam and Philadelphia: John Benjamins. 157-171.

Grant, L. (forthcoming), '"I mean, you know, I don't know" – comparing conversational use of "I don't know" between British and New Zealand speakers'.

Kärkkäinen, E. (2003), *Epistemic Stance in English Conversation.* Amsterdam and Philadelphia: John Benjamins.

Müller, S. (2005), *Discourse Markers in Native and Non-native English Discourse.* Amsterdam and Philadelphia: John Benjamins.

Östman, J.-O. (1981), *'You know': A Discourse-Functional Approach*. Amsterdam and Philadelphia: John Benjamins.

Östman, J.-O. (1995), 'Pragmatic particles twenty years after', in: B. Wårvik, S.-K. Tanskanen and R. Hiltunen (eds.) *Organization in Discourse*. Turku: University of Turku. 95-108.

Pichler, H. (unpublished), 'Form-function relations in discourse: the case of *I don't know*'. http://nwplinguistics.ncl.ac.uk/V12&13-HEIKE.PICHLER.pdf.

Rühlemann, C. (2007), *Conversation in Context. A Corpus-driven Approach.* London and New York: Continuum.

Sacks, H., E. Schegloff and G. Jefferson (1978), 'A simplest systematics for the organization of turn-taking in conversation' in: J. Schenkein (ed.) *Studies in the Organization of Conversational Interaction.* New York: Academic Press. 7-55.

Schegloff, E.A. (1996), 'Turn organization: one intersection of grammar and interaction', in: E. Ochs, E.A. Schegloff and S.A. Thompson (eds.) *Interaction and Grammar*. Cambridge: Cambridge University Press. 52-133.

Scheibman, J. (2000), '*I dunno*: a usage-based account of the phonological reduction of *don't* in American English conversation', *Journal of Pragmatics*, 37: 105-124.

Tsui, A.B.M. (1991), 'The pragmatic functions of *I don't know*', *Text*, 11(4): 607-622.

Appendix

Table 1a: The distribution of *I don't know* and *(I) dunno* in LINDSEI with regard to informant

	I don't know	*(I) dunno*		*I don't know*	*(I) dunno*
1	-	-	26	-	-
2	-	-	27	2	6
3	-	-	28	-	-
4	5	10	29	1	-
5	-	-	30	2	-
6	-	2	31	2	-
7	-	1	32	2	-
8	2	1	33	1	-
9	-	-	34	2	3
10	-	3	35	1	1
11	2	12	36	2	-
12	1	12	37	1	-
13	1	3	38	4	1
14	1	-	39	6	-
15	1	-	40	2	-
16	-	1	41	1	-
17	1	1	42	2	-
18	4	2	43	1	1
19	-	-	44	2	8
20	-	-	45	3	-
21	-	4	46	2	1
22	5	1	47	3	1
23	4	1	48	2	-
24	2	3	49	-	-
25	1	-	50	2	-

On the face of it: How recurrent phrases organize text

Magnus Levin and Hans Lindquist

Växjö University

Abstract

This study concerns the text-organizing functions of the recurrent phrases on the face of it*,* on its face *and* in (the) face of*. It is argued that the development of the pragmatic meanings is related to grammaticalization theory. It is demonstrated that the two former phrases often occur in constructions where they are followed by a hedge and a refutation (*on the face of it … this seems … But…*).* In (the) face of *mainly organizes text through its connection with negative evaluations (e.g.* in the face of opposition*). The new* Time *and COCA corpora compiled by Mark Davies are shown to be useful complements to the BNC. These corpora show that* on its face *is typical of American English, and that the article-less variant* in face of *is rare, and possibly decreasing in use.*

1. Introduction

1.1 Phraseology and text organization

As pointed out by Granger and Paquot (2008: 34), early work on recurrent phrases in spoken language often focused on their pragmatic, interactional function (cf. Aijmer 1996; Altenberg 1998). Coulmas (1981) said that such phrases are "expressions whose occurrence is tied to more or less standardized communication situations" (Coulmas 2-3 quoted in Wray 2002: 53) and in Nattinger and DeCarrico's pedagogical treatment of recurrent phrases in English two of the three main categories were "social interactions" and "discourse devices" (1992: 59-60). This interest in the pragmatic and discourse-organizing functions of recurrent phrases has been maintained in more recent research. For instance, Moon (1998) has a chapter on the function of FEIs (Fixed Expressions and Idioms) where, apart from the Informational (ideational) meaning, she lists Evaluative, Situational, Modalizing and Organizational meanings (Moon 1998: 217). Similarly, Granger and Paquot (2008: 41-45), basing their discussion on Burger (1998), make a tripartite division of the functions of phrases (or 'phrasemes' in their terminology) into Referential, Textual and Communicative. The textual function in its turn is exemplified by complex prepositions, complex conjunctions, linking adverbials and textual sentence stems. We will return to some of these categories in our discussion below.

From the psycholinguistic point of view, Wray (2002), in her comprehensive survey of formulaic language, also shows that formulaic sequences are used to signal discourse structure, noting that "[f]ormulaic discourse markers seem able to support both the speaker's and the hearer's processing simultaneously" (Wray 2002: 93).

Many of these recent studies on the pragmatic and discourse-organizing functions of recurrent phrases indicate that the meaning of the individual words in recurrent phrases is typically weakened, while the pragmatic meaning and discourse-organizing functions are strengthened (Stubbs 2007a: 165-166). This is a process parallel to established processes in grammaticalization where lexical items lose some of their semantic features while gaining new pragmatic features (cf. Hopper and Traugott 2003: 94). In a discussion of the relations between the study of phraseology and language change, Stubbs (2007b: 98) mentions the well-known fact that phrases containing body-part nouns often grammaticalize into adverbs and discourse markers. He goes on to exemplify this with phrases containing *face*:

> [T]he word *face* […] is used […] as a body term, and also as a place term (*the north face of the Eiger*). The 5-gram *in the face of the* is almost always used entirely abstractly, and is usually followed by a word denoting a problem. Similarly, *on the face of it* has the pragmatic function of introducing a potentially disputed interpretation. (2007b: 98-99)

Heine and Kuteva (2002) have noted that body-part nouns often occur as sources (starting points) of grammaticalization, and, from a different vantage point, cognitive linguists have shown that the world is often described metaphorically by means of references to basic domains of experience, like our bodies (Lakoff and Johnson 1980: 117; Goossens 1990; Gibbs et al. 2004): *the mouth of the river* (cf. Lindquist and Levin 2008), *the leg of the table* and *the face of the earth*. Moreover, phrases containing body nouns are used to describe the position in physical space taken up by human beings and physical entities: *at the foot of the mountain* (cf. Lindquist and Levin 2008), *on your right hand* (cf. Lindquist forthcoming), *at the back of the house* and *on the side of the road*. Likewise, a sequence of events can be described by means of the phrase *back to back* ('consecutively') (cf. Lindquist and Levin forthcoming). Through grammaticalization processes such expressions have developed into spatiotemporal terms (cf. Claudi and Heine 1986 and the discussion in Hopper and Traugott 2003: 85). Similarly, discourse and texts can be organized by means of body phrases: *on the one hand* (cf. Lindquist forthcoming), *at the foot of the page* (cf. Lindquist and Levin 2008).

In spite of various forays by corpus linguists into the area of discourse analysis over the years, there is still some truth in Lee's (2008) claim (based on a small survey in 2005) that "very little discourse analytic work [has] been done by corpus linguists" (86). And in the field of pragmatics, there has been a certain

reluctance to use corpora in spite of the fact that pragmatics is the study of language in use (cf. Pons Bordería 2008: 1353-1354). In the present paper, we will investigate a number of phrases containing the body noun *face* which have developed pragmatic, text-organizing functions. In doing so, we want to combine the insights and methods of traditional phraseology and the fine-grained qualitative analysis of discourse analysis with a corpus-linguistic approach which entails a certain amount of data-drivenness and open-mindedness as to what constitutes a phrase or formulaic sequence in the first place.

1.2 Aim and scope

This study will provide new insights into the pragmatic and text-organizing functions of recurrent body-noun phrases whose origins and current usage are based on the embodied nature of human language. The aims of the study are to:

- describe and analyze the most frequent recurrent text-organizing phrases containing the noun *face*;
- analyze the sentence and discourse contexts in which these phrases are used;
- search for traces of the diachronic development of these phrases;
- relate the development of the pragmatic meanings of these phrases to the concept of grammaticalization.

Two complementary, methodological aims are (i) to develop ways of relating quantitative corpus data to qualitative, fine-grained analyses of the phrases in their discourse environment and (ii) to see if this is doable using the web interface of a number of corpora supplied by Brigham Young University (Davies 2004, 2007-, 2008-) in combination with the database *Phrases in English* (PIE) and its interface, supplied by Fletcher (2003/2004).

2. Material and method

The British English material for this study comes from the *British National Corpus* (BNC), which consists of 100 million words (90 million written; 10 million spoken). This was compared with American English data from the recently completed Brigham Young *Corpus of Contemporary American English*, (COCA) (Davies 2008-). This corpus is made up of more than 360 million words of American English (as of spring 2008) from 1990-2007 downloaded from books, journals, magazines and transcribed radio and television broadcasts on the Internet. In order to access diachronic material, we also investigate the *Time* corpus (Davies 2007-), which consists of all issues of *Time* magazine between 1923 and 2006.

For the main part of the study, recurrent phrases containing the body noun *face* were extracted from the BNC via Fletcher's PIE database (Fletcher 2003/2004) which contains n-grams (identical strings of words) with a length between two and eight words occurring at least three times in the corpus. We repeated the searches with *face* in different positions from 8-grams all the way

down to 2-grams. Those n-grams that appeared to have discourse-organizing functions were investigated further. For instance, figure 1 illustrates how the search for 5-grams was carried out with the word *face* in different positions, and exemplifies the kind of strings that were found. Frequency figures are given in brackets.

face + + + +	*face* to face with the (58)
+ face + + +	the *fac*e of the earth (66)
+ + face + +	on the *face* of it (257)
+ + + face +	to change the *face* of (14)
+ + + + face	the look on his *face* (31)

Figure 1: 5-grams with *face*

In this case, only *on the face of it* was taken into consideration, while the others were discarded as not being involved in discourse organizing. With this method, the researcher retrieves the most central and typical recurrent phrases without having to rely on preconceived lists based on introspection or dictionaries (for further discussion, see Stubbs 2007a, 2007b; Lindquist and Levin 2008). When we had sifted through the recurrent word combinations found by PIE, we arrived at the following phrases that organize discourse: *on the face of it*, *on its face* and *in (the) face of*. Other *face*-phrases, which can serve as a basis for further studies, are more marginally used as text-organizers. A couple of these are seen in examples (1) and (2), where *taken at face value* and the verbless *at face value* occur sentence-initially (italics added in all examples).[1]

(1) These were the strongest statements yet issued in the month-long battle (TIME, Nov. 27), directed from Washington, to raise U.S. production. *Taken at face value*, they *looked like* real cause for alarm. *But* the simple fact was: the U.S. has not yet lost a battle for want of supplies. (*Time,* 1944/12/18)

(2) *At face value*, The Magic Flute is a musical fairy-tale with a rather silly plot, cooked up by Schikaneder from a variety of sources. The story concerns the efforts of Prince Tamino to rescue Pamina, daughter of the Queen of the Night […]. So far, The Magic Flute *appears* to fit neatly into the Viennese Singspiel tradition, with its emphasis on simple music, and dramatic effects (monsters, choruses of dancing slaves, etc.) drawn from pantomime. *But* scratch the surface, and an astonishing mystery is revealed. (BNC: CEW)

Like the prototypical instances of *on the face of it* and *on its face*, (1) and (2) contain hedges ("looked like") and refutations ("But…") (see below for further discussion).

3. Results

Face is among the 75 most frequent nouns in the BNC (Leech et al. 2001) and is the sixth most common body noun after *hand*, *head*, *side*, *eye* and *body*. There are more than 26,000 instances of *face* occurring as nouns in the BNC, more than 16,000 instances in *Time*, and 100,000 instances in COCA. It has been noted by, among others, Stubbs (2007b: 100) and Lindquist and Levin (2008: 144-145) that frequent nouns often occur in frequent phrases and that in such recurrent word combinations they tend to express meanings towards the non-literal end of the meaning spectrum. This is also the case with *face*, which only occurs in its most literal meaning 'the front part of the head' (as in *the warm evening air brushed her face*) in about a third of the cases. The results of an analysis of 300 random instances are shown in table 1 (based on 100 tokens each from the BNC, COCA and *Time*).

Table 1: 300 random tokens of the noun *face* from the BNC, COCA and *Time*

	Literal	Metonymical			Metaphorical		Other	Total
		Facial expression	*face to face*	Person	Surface	Text-organizer		
BNC	37	35	7	4	9	3	5	100
COCA	49	16	8	14	5	6	2	100
Time	30	12	9	5	23	19	2	100
Total	116	63	24	23	37	28	9	300

The categories in the table are partly overlapping in that, for instance, 'facial expression' also includes literal references to the face. As indicated above, the figures for literal meanings are similar to those found in previous studies, such as Lindquist and Levin (2008: 147), where around a third of the instances of the lemma FOOT and around half of those of MOUTH referred to actual body parts. There were three main groups of metonymic meanings with *face*: 'facial expression' (*his big red face shone with kindness*), the high-frequency positional phrase *face to face* and 'person' (*he is a familiar face*). Metaphorical meanings comprise meanings where *face* refers to various surfaces (*the North Face*) and to phrases used to organize text, such as *on the face of it*. When sequences like *on its face* and *in the face of* occurred in compositional phrases such as *a horse with a big smile on its face* they were classified as literal. According to Wray (2002: 262-264), sequences can be both stored holistically in the mental lexicon and created compositionally in the communication situation. It is noteworthy that approximately ten per cent of the instances of the word *face* in the three corpora occur in metaphorical text-organizing phrases.

It seems that there are some differences in the distribution of the various kinds of phrases between the broadly sampled BNC and COCA, and the more specialized *Time* corpus. For example, literal meanings are somewhat more

frequent in COCA than in the two other corpora and metaphors are more frequent in *Time*, while, because of the large proportion of fiction in the BNC, there is a large proportion of descriptions of facial expressions in that corpus. In addition, text-organizing phrases with *face* seem more common in *Time*, largely due to the high incidence of *in (the) face of*. This phrase is based on the "confrontational" connotations of *face*, while *on the face of it* and *on its face* are based on the contrast between a surface and what may lie under it. We will begin by discussing the latter two phrases in 3.1.

3.1 *On the face of it* and *on its face*

The text-organizing uses of *on the face of it* and *on its face* appear to be of roughly similar age: the first attestation in the *Oxford English Dictionary* (OED) for *on the face of it* is from 1879 (*The inter-relationship of these two subjects may not seem* on the face of it *very clear, but inter-relationships of customs very rarely are* (OED s.v. interrelationship)), and the first attestation for *on its face* is American English from *The New York Times* from 1904 (*The police are now forced to take what appears* on its face *to be the veriest pipe story and run it down* (OED s.v. pipe story)).[2] In spite of the early attestation of *on its face*, Quirk et al. (1985: 362) star it in the example *On the face of it / *On its face it seems a good idea*. This judgement of ungrammaticality will be discussed below.

Before moving on to the more typical instances where *on the face of it* and *on its face* express a contrast, it should be noted that they occasionally indicate a result according to expectation, as illustrated in (3), which contains the adverb *indeed* stating that things are really what they seem on the surface.

(3) *On the face of it*, it is a business ripe for computerization, and, *indeed*, a great deal of work is done with computers to analyse media research data and make the construction of media schedules more effective. (BNC: F9D)

More typically, however, the phrases are followed by hedges, such as *seem*, which strengthen their hedging function, as illustrated in (4).

(4) *On the face of it*, that *seems* neither more nor less than Saddam has offered before. (COCA: ABC Nightline)

Moreover, the text-organizing function is perhaps most clearly seen when the phrases are followed by explicit contradictions, as in (5):

(5) *On the face of it*, they were an utter failure. *But* that is too simple. (BNC: ACS)

Finally, the prototypical pattern is seen in (6), which contains both a hedge ("would appear") and a refutation containing the explicit contrast ("But one has to look beyond the surface").

(6) *On its face*, the U.S. experience with refrigerator energy efficiency *would appear* to satisfy even the most demanding sustainability critics. A 75-percent reduction in refrigerator electricity requirements is substantial, and the fact that ozone-depleting chemicals (chlorofluorocarbons) have been practically eliminated from U.S. refrigerators is praiseworthy indeed. *But* one has to look beyond the surface to see where problems arise. (COCA: Environment)

A large proportion of the instances of *on the face of it* and *on its face* co-occur with hedges and are followed by explicit contradictions to the state of affairs that superficially holds true. This construction thus allows writers to downplay or background the surface appearance of a phenomenon. By looking at the extended context, we have thus moved the focus from n-grams generated by the search engine to discontinuous text-organizing constructions that may extend over several sentences.

Before we turn to the analysis of the text-organizational functions of the phrases, we will consider the overall distributions in the three corpora and the diachronic pattern in the *Time* corpus. The distributions of *on the face of it* and *on its face* are given in figure 2. The frequencies per ten million words are given above, and the real numbers below, the bars.

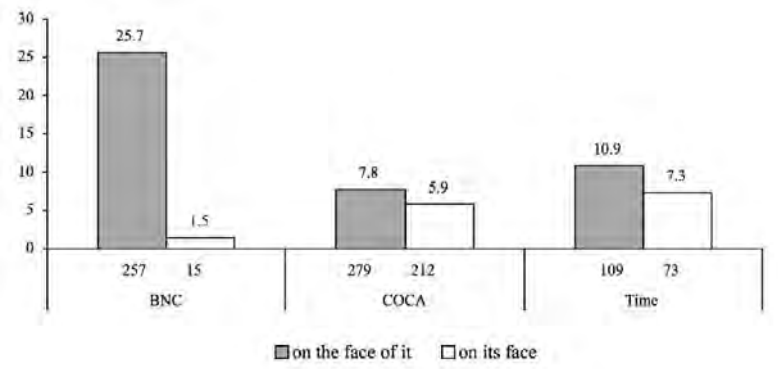

Figure 2: *On the face of it* and *on its face* in the BNC, COCA and *Time* (frequency per ten million words)

As seen in figure 2, *on the face of it* is the more frequent alternative in all three corpora, although the phrases are almost equally frequent in the American English material. There are noteworthy differences between the varieties, however. The preference for *on the face of it* is stronger in British English (a chi-squared test indicates that the differences between the BNC and the American English corpora are significant ($p \leq 0.01$)). The low frequency of *on its face* in British English probably explains Quirk et al.'s (1985: 362) dislike for this variant.

Of course, the results only tell us that these British and American corpora are different. It is a bit precarious to generalize about the whole language from comparisons between corpora with different contents such as the BNC, COCA and *Time*. However, as seen below, *on its face* is restricted to legal texts in the BNC while it has a much wider distribution in COCA, so the differences between the corpora are in all likelihood fairly accurate reflections of the variation between British English and American English.

Next we shall see that combined evidence from text-type variation in COCA and the diachronic pattern in *Time* suggest that *on its face* is a well-established alternative in formal American English, where it may even be on the increase. To begin with, *on its face* is most frequent (eleven instances/ten million words) in the sub-corpus of COCA containing academic text, which, according to Mair (1998: 155) and Hundt and Mair (1999), is a "slow" or "uptight" genre that is conservative when it comes to accepting innovations. This suggests that *on its face* has been established as an alternative in American English for a long time, although it does not indicate anything more specific about its diachronic development. The diachronic *Time* corpus is therefore a useful complement to the more synchronic COCA. In figure 3, we present the findings for *on the face of it* and *on its face* in *Time*. The frequencies per ten million words are presented above, and the real numbers below, the bars.

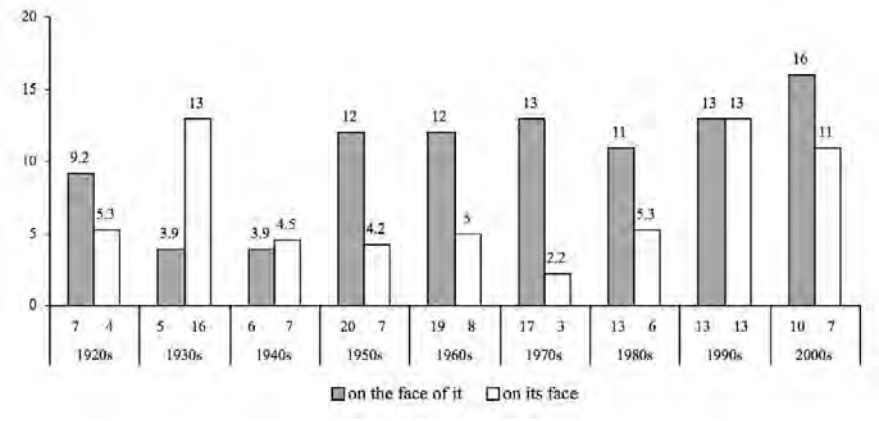

Figure 3: *On the face of it* and *on its face* in the *Time* corpus (frequency per ten million words)

There are no unequivocal trends, but the material at least suggests that neither of the alternatives is decreasing in use. Instead, although it should be stressed that the numbers are small, there is some suggestive evidence that both are on the increase: the three lowest frequencies for *on the face of it* occurred in the first three decades, while the highest figure is for the last decade; for *on its face*, the two highest frequencies per ten million words occurred in the last two decades, apart from a peak in the 1930s material.[3]

The potential increase of both phrases is interesting in view of Sapir's (1971: 165) claim that analytic *of it* is beginning to oust synthetic *its*. Although it is highly doubtful that *its* is disappearing, it is nevertheless striking that both alternatives appear to be on the increase here, since this means that the variability within a metaphorical adverbial is becoming established. Hudson (1998: 33-36) proposes that multi-word adverbs in general are not analyzable and that this is one of the reasons why they tend to resist variation. Judging from this, the variation between *on the face of it* and *on its face* would be exceptional, but, as will be seen in 3.2, *in the face of* also has a variant form, *in face of*.

A more fine-grained analysis of the data shows again that *on its face* is particularly frequent in legal language: all 15 instances in the BNC and around a quarter of the instances in COCA occurred in legal contexts (where it collocates with words like *(un-)constitutional*, *legislation* and *statute* (as in (14) below)). The difference in frequency between American English and British English can thus mainly be explained by *on its face* having spread to non-legal uses as a text-organizing device in American English.

Interestingly, four of the legal instances of *on its face* in COCA were contrasted with the phrase *as applied* (as in, e.g., *Eastern claimed that the Coal Act, either* on its face *or as applied, violated substantive due process*) which indicates that some of these tokens have a specialized legal sense and only marginally serve as text-organizers. However, hedges are about equally frequent with *on its face* in legal contexts (as in, e.g., *a law which appeared* on its face *to discriminate among religions*) as in other contexts, while explicit contradictions are markedly rarer (see further discussion below).

The overall distributions of hedges and contradictions with the two phrases in BNC and COCA are presented in table 2, which shows the total number of tokens of each phrase, tokens with only hedges, tokens with only contradictions and tokens with both hedges and contradictions.

Table 2: The co-occurrence of *on the face of it* and *on its face* with text-organizing features in the BNC and COCA

	on the face of it				on its face			
	Total	Hedge	Contra-diction	Both	Total	Hedge	Contra-diction	Both
BNC	257	64 (25%)	68 (26%)	59 (23%)	15	2 (13%)	0 (0%)	2 (13%)
COCA	279	89 (32%)	37 (13%)	62 (22%)	212	58 (27%)	35 (17%)	33 (16%)

We will first discuss instances without hedges or contradictions, then instances with only contradictions, then instances with only hedges, and finally instances with both hedges and contradictions.

As seen in the table, between a quarter and 40 percent of all instances (except for *on its face* in the BNC where the numbers are very low) contain neither

hedges nor explicit contradictions, as in (3) above and (7) below. In (7) the phrase is parenthetically inserted together with an epistemic marker of high likelihood. Thus *on the face of it* and *on its face* may not only express a counter-expectation (see below), but occasionally also indicate a result according to expectation. The phrases can therefore be argued to have more than one text-organizing function.

(7) Listen to me Jenny. If I only give you one piece of advice in life – which is highly likely *on the face of it* – it's stay away from that man. (BNC: A0F)

Contradictions are typically overtly introduced by *but*, as in (8), *however*, as in (9), or *yet*, as in (10).[4] However, a number of instances do not involve connectors, as illustrated in (11).

(8) "*On its face*, it's a pretty cool plan," he says. "*But* with our margins, it would be easy to give back a quarter's worth of profitability due to one startup company that leaves you holding the bag on $15 million worth of materials". (COCA: Fortune)
(9) The control layout is on a par with ART's tried and tested formula, and *on the face of it* should be fairly simple to understand. *However*, as a lot of the controls serve two functions, some initial confusion can arise. (BNC: C9M)
(10) *On the face of it*, lying around in a deep torpor, exposed to attack, does not *sound* like a smart move. *Yet* we and other mammals can not do without our sleep. (COCA: FantasySciFi)
(11) Working among books was, *on the face of it*, a ladylike occupation, Mrs Broome had thought, and one that would bring her daughter into contact with a refined, intellectual type of person. She had never seen Ianthe handing out books to the ill-mannered grubby students and cranks of all ages who frequented the library of political and sociological books where she worked. (BNC: HA4)

On the face of it and *on its face* often co-occur with hedges which strengthen the counter-expectation or anticipation of the upcoming refutation that has already been created (for the notion of counter-expectation, cf. Heine, Claudi and Hünnemeyer 1991: 192-204). This is exemplified in (12), which incidentally also contains the phrase *in the face of*, to (15) below. By far the most frequent hedges are *seem* and *appear*.[5] *Seem* is the most common, occurring 159 times overall, while *appear* occurs 52 times. Thus two types of hedges account for more than half the instances in the entire material. Stubbs (forthcoming) argues that it is typically the case in variable phrases that a few alternatives are used in the majority of the instances, and that some choices can be considered to be canonical.

(12) *On the face of it* such an action *appears* to be a simple case of cowardice in the face of the enemy. *However*, I was not willing to order the setting up of a court-martial before I had heard your version of what took place that morning. (BNC: K8T)
(13) *On the face of it*, the Government *appears* to be putting more money into education through higher pay for teachers, *but in reality* looks set to claw it back by making education authorities find that cash from within their existing budgets. (BNC: K54)
(14) *On its face*, the statute *might seem* to block any consideration of cost and, indeed, to require regulations that would reduce risks to zero, especially because for many toxic substances safe thresholds simply do not exist. (COCA: Michigan Law Review)
(15) *On its face*, this new therapeutic sex pedagogy does *not seem* all that therapeutic or all that new. (COCA: Atlantic)[6]

Hyland (1998: 9) notes that some linguists include hedges among text-organizing elements since they "explicitly organise the discourse, engage the audience and signal the writer's attitude". This also applies to the constructions consisting of *on the face of it / on its face* followed by a hedge and an explicit contradiction. The *face*-phrase and the additional hedge indicate the writer's reservation regarding the truth value: what is discussed at first is only the surface impression of an issue and readers should be prepared for a refutation. Such explicit text organization can be argued to be slightly redundant since sentences without either of the two initial hedges work nicely as well, as illustrated in the shortened versions of (12) given as (12a) and (12b) below:

(12a) Such an action *appears* to be a simple case of cowardice in the face of the enemy. *However*, I was not willing to order the setting up of a court-martial before I had heard your version of what took place that morning.
(12b) *On the face of it* such an action is a simple case of cowardice in the face of the enemy. *However*, I was not willing to order the setting up of a court-martial before I had heard your version of what took place that morning.

By drawing attention to the surface, a speaker or writer typically implies that there is a difference between what a phenomenon seems to be like and what it really is like.

In their discussion of *indeed*, *actually* and *in fact*, Traugott and Dasher note that later pragmatic meanings "are to various degrees polysemous with other, earlier uses of the same lexeme/construction" (2005: 157). Layering of this kind is not uncommon in our material, so that in some cases apparent facts are indeed what they appear to be on the face of it, as illustrated in (3) above. Such instances relate both to the epistemic status of the utterances and to text-organization, since they serve to stress that things are really what they seem and use the surface appearance of a phenomenon as a starting point for the discussion of it.

The different functions of *face*-phrases are summarized in the chain in figure 4. This chain instantiates the grammaticalization chain proposed by Traugott (1982), which has since been seen in a number of variants. This particular chain starts with the original propositional sense 'front part of the head' in (i), a metaphorical extension to 'surface' in (ii), and finally in (iii) there is the entrenchment of two recurrent, semantically opaque phrases that occur in text-organizing constructions.

(i)	literal *face*	>	(ii)	surface *face of sth.*	>	(iii)	text-organizer *on the face of it* (*on its face*)

Figure 4: From propositional to textual meaning

Because *on the face of it* is (virtually) exclusively used as a text-organizer, this phrase is arguably more grammaticalized than *on its face*. The development of *on its face*, on the other hand, can be argued to instantiate the layering of meaning. The sequence *on its face* can be found in reference to (i) faces and (ii) surfaces and also in the function of a text-organizer (iii), usually in a larger construction. In the latter stage, the original meaning of the noun *face* has been lost and the recurrent phrase has only pragmatic or text-organizing functions, which is a typical development path of multi-word units, according to Stubbs (2007a: 165-166).

In conclusion, *on the face of it* and *on its face* are both based on a metaphorical link between the face and a surface, and from this to the surface impression of something, which in turn often implies a deceptive appearance. Since bodily-based phrases are used to express textual meanings, these phrases can be seen as typical examples of the embodied nature of language. Furthermore, both types of phrases typically form parts of bigger constructions consisting of a hedge strengthening the counter-expectation and an overt contradiction of the apparent facts (e.g. "On the face of it … appears … However …", as in (12) above). Finally, both text-organizational and propositional meanings co-exist (layering), so that at least *on its face* in some instances has kept its older propositional meaning. The open-endedness and variability in use of the two phrases, and the fact that in actual discourse they occur in a number of lexico-grammatical patterns, make it reasonable to categorize them loosely as "textual sentence stems", using Pawley and Syder's (1983) term which has recently been recycled by Granger and Paquot (2008: 42) in their classification of textual phrasemes.

The next section concerns a phrase which has developed in a slightly different manner.

3.2 *In (the) face of*

The OED (s.v. face, n. 4.c.) gives three meanings for *in (the) face of*: "(a) in front of, directly opposite to; (b) face to face with, when confronted with; (c) in

defiance of, in direct opposition to, notwithstanding". There is a clear semantic development from the concrete meaning in (a) to the abstract meanings in (b) as illustrated in *In the face of bad example, the best of precepts are of but little avail* (1871) (s.v. face, 4.b.) and in (c) as illustrated in *They now assert here, in the face of facts, that the cholera has ceased* (1837) (s.v. face, 4.c.). It can be argued from the various senses in the OED that *in (the) face of* has developed into a complex preposition.

In (the) face of, which is mentioned briefly as a text-organizing phrase by Moon (1998: 217), thus has its metaphorical origin in the confrontational connotations of *face*.[7] Such connotations are also present in other *face*-phrases, such as *in your face* ('direct and often shocking') and *come face to face* with sth./sb. ('confront'). The phrase *in (the) face of* usually refers to a confrontation with an obstacle of some sort, such as opposition or adversity, and is therefore clearly associated with an element of negative evaluation. Stubbs (2007a: 166) argues that such evaluations are part of text-organization since "evaluating something focuses attention on it, contrasts it with something else, and emphasizes its importance in the text".

As indicated above, *in the face of* sometimes occurs without the definite article (as is typical in the development of complex prepositions): *in face of*. The share of short forms in the three corpora is as follows: BNC four percent (62/1461),[8] COCA one percent 66/5300) and *Time* two percent (33/1900). Grammaticalization is often considered to be the result of conversational implicatures which have occurred frequently over time. While this could be the case for the frequent long form *in the face of*, some other explanation needs to be sought for the infrequent short form *in face of*. Hoffmann (2004: 193-195) has presented an interesting hypothesis, viz. that infrequent sequences may grammaticalize by analogy with sequences which have similar form and a similar function (in this case text-organizing). This is likely to be the case for *in face of*. From a diachronic point of view, one would expect the shorter form to be a later development, replacing the earlier form.[9] However, currently *in face of* appears to be a formal alternative that is possibly becoming obsolete: it occurs mainly in academic texts in the BNC and COCA, and the last attestation in *Time* is from the 1970s.

The two main meanings of *in (the) face of* are sometimes hard to distinguish. The more frequent one can be paraphrased as 'when confronted with' (OED), as in (16) and (17). The less frequent meaning can be paraphrased as 'despite', as exemplified in (18) and (19), the latter of which actually contains an explicit *despite*-phrase.

(16) This can also happen when a doctor experiences discomfort *in the face of* death, or an inability to come to terms with his own helplessness. (BNC: ASK)

(17) *In face of* this opposition, the U. S. delegation, last week, quit Geneva, entrained for Paris en route for Washington, washed its hands of the conference. (*Time*, 1925/02/16)

(18) We are maintaining our high promotional investment in the carpet sector and, *in the face of* strong competition, the third quarter has started satisfactorily. (BNC: HRY)
(19) the referendum had gone ahead *in the face of* claims by the federal authorities that it was unconstitutional, and *despite* threats of economic sanctions. (BNC: HL2)

The most frequent objects occurring in the frame *in the face of* N all unequivocally express negative connotations: *death*[10] (BNC 11, COCA 66, *Time* 19), *adversity* (BNC 13, COCA 68, *Time* 10), *opposition* (BNC 16, COCA 21, *Time* 7); while the most frequent adjectives (*mounting* (BNC 11, COCA 35, *Time* 18), *overwhelming* (BNC 7, COCA 29, *Time* 16), *growing* (BNC 11, COCA 28, *Time* 10)) are implicitly negative since they usually collocate with nouns with negative connotations (e.g. *mounting criticism*). Preceding collocates are much rarer, but at least the two most frequent adjectives *helpless* (BNC 14, COCA 26, *Time* 6) (as in (20)) and *silent* (BNC 6, COCA 19, *Time* 8) (as in (21)) are clear instances of negative evaluations.

(20) Americans often feel *helpless in the face of* crime, and so they sometimes respond by overreacting. (COCA: Cosmopolitan)
(21) Our problems and moral failings are not from unnecessarily publicizing our shortcomings, but of remaining *silent in the face of* horrendous crimes being committed daily against the very existence of the Umma. (COCA: IntlAffairs)

We have opted here for the more theory-neutral terms 'evaluation' and 'connotations' (Levin and Lindquist 2007) rather than, for instance, 'semantic prosody' (Louw 1993; Partington 2004; Sinclair 2004) or 'discourse prosody' (Stubbs 2007a: 178-179). By choosing 'evaluation' and 'connotations' as our terms we avoid some problems, such as having to establish to what extent meanings are explicitly encoded in the words themselves and to what extent they acquire their meanings from their contexts (for further discussion of the problems of semantic prosody, see Whitsitt 2005; Levin and Lindquist 2007).

The meanings of *in (the) face of* can be difficult to distinguish even in rather extended concordance contexts, as illustrated in (22), where either 'despite' or 'when confronted with' seems applicable. Other meanings than those found in dictionaries also occur, as seen in (23), where *in the face of* seems to mean something like 'instead of' and in (24), which we came across in a football magazine, where it might be paraphrased by 'to the benefit of'.

(22) But starting the Reformation was only half the battle: turning its principles into practicable policies and maintaining momentum *in the face of* opposition from determined vested interests was much more difficult. (BNC: ABA)

(23) In the past the CEGB had been criticised for its lack of foresight in building high operating cost oil-fired stations and expensive nuclear stations *in the face of* cheaply fuelled coal-fired stations. (BNC: AT8)

(24) For years football was all but ignored *in the face of* other sports considered more likely to reap international successes. (*When Saturday Comes*, 254: 38)

Although there are differences between *on the face of it / on its face* and *in (the) face of* in that the former are classified as 'textual sentence stems' and the latter as a complex preposition associated with negative evaluation, and in that they are based on different aspects of the *face* metaphor, there are also many important similarities. Both *on the face of it / on its face* and *in (the) face of* exemplify in their origins a movement from propositional to pragmatic or text-organizing functions where the basic meaning of the noun *face* has been weakened. Furthermore, both kinds of phrases are polysemous. *In (the) face of* expresses various meanings, the most frequent of which are 'when confronted with' and 'despite'; *on the face of it* and *on its face* typically express counter-expectation, but they also occasionally express a result according to expectation, and in addition *on its face* retains a specialized legal meaning.

4. Summary and conclusions

A preliminary survey of the uses of the individual word *face* in the corpora showed that in as many as two-thirds of the occurrences there is no reference at all to an actual face; instead *face* is used with a number of figurative meanings and in a number of different phrases or formulaic sequences. Ten percent of the tokens of *face* occurred in phrases with pragmatic and text-organizing functions. We found two main types of text-organizing sequences, each based on a different metaphorical link: *on the face of it / on its face* (surface) and *in* (*the*) *face of* (confrontation).

On the face of it and *on its face* typically form part of bigger constructions containing a hedge (e.g. *seem*) and an overt contrast to what appears to be the case, introduced by, for example, *but* or *however*, usually one or several sentences later in the text. Due to this open-endedness we have classified *on the face of it / on its face* as 'textual sentence stems'. The two variants have distinct distribution patterns over regional varieties, so that *on its face* is relatively rare in British English, where it is mainly confined to legal registers, while in American English it competes rather successfully with *on the face of it* in all registers. In our diachronic distribution data from the *Time* corpus, there were no clear trends except an indication that both *on the face of it* and *on its face* are becoming more frequent in American English.

As regards the string *in (the) face of*, it can be argued that it has grammaticalized into a complex preposition. One sign of this is the existence of the rare form without the definite article, *in face of*. However, our corpora give no

indication that the shorter form is on the increase, which might have been expected. While *on the face of it / on its face* forms parts of constructions that typically contrast the surface appearance with the real state of affairs, *in (the) face of* organizes text by expressing negative evaluation.

The results show that the body noun *face* occurs frequently in a number of recurring formulaic sequences which are used by speakers and writers to organize discourse and to express evaluation. This supports both the claims of cognitive linguistics that language is embodied and the claims from grammaticalization studies that phrases can develop along a path from propositional to subjective and textual meanings.

From a corpus methodological point of view, our study shows that COCA and the *Time* corpus provide a wealth of relevant material for the synchronic and diachronic study of American English, which to some extent is comparable to material from the BNC. However, the COCA interface has its limitations in large-scale studies of text-organization because the procedure of expanding the context proved to be rather cumbersome, with each line having to be clicked on individually to retrieve more context. A further disturbing factor was that using the Brigham Young University interface to search the BNC we did not retrieve any hits for some strings although the same strings could be found through the Zurich BNC-Web interface.

Finally, we believe that we have shown that a combination of corpus methodology, where phrases or formulaic sequences are identified and retrieved semi-automatically, and close discourse analytic scrutiny of the textual environment of individual tokens beyond the sentence borders, is a fruitful mode of operation for pragmatic studies of textual organization.

Notes

1 *Taken at face value* occurs 26 times in the material (4 in the BNC, 12 in COCA, 10 in *Time*) and *at face value* alone occurs 19 times (6 in the BNC, 11 in COCA and 2 in *Time*).

2 Note that the two earliest attestations both contain hedges (*may not seem / appears*) and that *on the face of it* also co-occurs with a refutation (*but*).

3 We scrutinized the individual examples but found no likely explanation for this high frequency in the 1930s, so it may well be due to chance.

4 The most frequent connectors co-occurring with *on the face of it* were *but* (BNC 61, COCA 59), *however* (BNC 27, COCA 11) and *yet* (BNC 10, COCA 11). The numbers for *on its face* were as follows: *but* (BNC 1, COCA 36), *however* (BNC 0, COCA 4) and *yet* (BNC 1, COCA 6). Thus a restricted set of words (*but*, *however*, *yet*) are used to express contradiction in these constructions.

5 These verbs occasionally co-occur with *may* and *might*, which, according to Hyland (1998: 116), are sometimes considered to be the prototypical hedges. In connection with *on the face of it* and *on its face*, however, they are much rarer than *seem* and *appear*.

6 Furthermore, (15) also illustrates the fairly frequent use of negations with the two phrases. For instance, *on its face* occurs with negations in 23 percent (49 of 212) of the tokens in COCA.

7 The phrase FLY *in the face of* ('to act in direct opposition to' (OED)) has not been included in the discussion, since it has a different pragmatic function than most instances of *in (the) face of*. Its text-organizing function is marginal, and FLY *in the face of* mainly expresses dislike of the act of opposing or contradicting something, as in *I challenge Dr McNab to justify his so-called remedies which* fly in the face of *all that's known about the pathology of this disease*. (BNC; EFW)

8 *In face of* had to be searched via the Zurich BNC interface (http://escorp.unizh.ch). The Brigham Young interface seems to be based on an older version of the BNC, from which, for some reason, it is not possible to retrieve this 3-gram.

9 This expected development is illustrated by the article-less complex preposition *in light of* which is replacing *in the light of* in the *Time* corpus.

10 But Cook (2008: 314-315) points out that even *death* can sometimes be seen as positive, e.g. for a suicide bomber.

Primary Sources

British National Corpus (BNC) 1995, Accessed via Davies (2004), W. Fletcher (<http://pie.usna.edu>) and BNC-Web.

Davies, M. (2004), BYU-BNC: *The British National Corpus*. Available online at <http://corpus.byu.edu/bnc>

Davies, M. (2007-), *Time Magazine Corpus* (100 million words, 1920s-2000s). Available online at <http://corpus.byu.edu/time>

Davies, M. (2008-), *Corpus of Contemporary American English* (360 million words, 1990-present). Available online at <http://www.americancorpus.org>

References

Aijmer, K. (1996), *Conversational Routines in English*. London: Longman.

Altenberg, B. (1998), 'On the phraseology of spoken English: the evidence of recurrent word-combinations', in: A. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Clarendon Press. 101-122.

Burger, H. (1998), *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmitt.

Claudi, U. and B. Heine (1986), 'On the metaphorical base of language', *Studies in Language*, 10: 297-335.

Cook, G. (2008), 'Hocus pocus or God's truth: the dual identity of Michael Stubbs', in: A. Gerbig and O. Mason (eds.) *Language, People, Numbers. Corpus Linguistics and Society*. Amsterdam: Rodopi. 305-327.

Coulmas, F. (1981), 'Introduction: conversational routine', in: F. Coulmas (ed.) *Conversational Routine*. The Hague: Mouton de Gruyter. 1-17.

Gibbs, W.R. Jr., P.L. Costa Lima and E. Francozo (2004), 'Metaphor is grounded in embodied experience', *Journal of Pragmatics*, 36: 1189-1210.

Goossens, L. (1990), 'Metaphtonomy: the interaction of metaphor and metonymy in expressions for linguistic action', *Cognitive Linguistics,* 1: 323-340.

Granger, S. and M. Paquot (2008), 'Disentangling the phraseological web', in: S. Granger and F. Meunier (eds.) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins. 27-49.

Heine, B., U. Claudi and F. Hünnemeyer (1991), *Grammaticalization: A Conceptual Framework.* Chicago: Chicago University Press.

Heine, B. and T. Kuteva (2002), *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hoffmann, S. (2004), 'Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition', in: H. Lindquist and C. Mair (eds.) *Corpus Approaches to Grammaticalization in English*. Amsterdam: John Benjamins. 171-210.

Hopper, P.J. and E.C. Traugott (2003) [1993], *Grammaticalization*. Second edition. Cambridge: Cambridge University Press.

Hudson, J. (1998), *Perspectives on Fixedness: Applied and Theoretical*. Lund: Lund University Press.

Hundt, M. and C. Mair (1999), '"Agile" and "uptight" genres: the corpus-based approach to language change in progress', *International Journal of Corpus Linguistics,* 4(2): 221-241.

Hyland, K. (1998), *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Lakoff, G. and M. Johnson (1980), *Metaphors We Live By*. Chicago: Chicago University Press.

Lee, D.Y.W. (2008), 'Corpora and discourse analysis: new ways of doing old things', in: V.K. Bhatia, J. Flowerdew and R.H. Jones (eds.) *Advances in Discourse Studies*. London: Routledge. 86-99.

Leech, G., P. Rayson and A. Wilson (2001), *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Levin, M. and H. Lindquist (2007), 'Sticking one's nose in the data. Evaluation in phraseological sequences with *nose*', *ICAME Journal,* 31: 87-110.

Lindquist, H. (forthcoming), 'A corpus study of lexicalized formulaic sequences with preposition + *hand*', in: B. Corrigan, E. Moravcsik, H. Ouali and K. Wheatley (eds.) *Formulaic Language: Volume 1. Distribution and Historical Change.* Amsterdam: John Benjamins.

Lindquist, H. and M. Levin (2008), 'FOOT and MOUTH. The phrasal patterns of two frequent nouns', in: S. Granger and F. Meunier (eds.) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins. 143-158.

Lindquist, H. and M. Levin (forthcoming), 'The grammatical properties of recurrent phrases with body part nouns: the N_1 to N_1 pattern', in: U. Römer and R. Schulze (eds.) *Exploring the Lexis-grammar Interface*. Amsterdam: John Benjamins.

Louw, B. (1993), 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in: M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins. 157-176.

Mair, C. (1998), 'Corpora and the study of the major varieties of English: issues and results', in: H. Lindquist, S. Klintborg, M. Levin and M. Estling (eds.) *The Major Varieties of English. Papers from MAVEN 97*. Växjö: Acta Wexionensia. 139-157.

Moon, R. (1998), *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Clarendon.

Nattinger, J.R. and J.S. DeCarrico (1992), *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Oxford English Dictionary, www.oed.com.

Partington, A. (2004), '"Utterly content in each other's company". Semantic prosody and semantic preference', *International Journal of Corpus Linguistics,* 9(1): 131-156.

Pawley, A. and F.H. Syder (1983), 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency', in: J.C. Richards and R.W. Schmidt (eds.) *Language and Communication*. London: Longman. 191-226.

Pons Bordería, S. (2008), 'Introduction to the special issue on empirical data and pragmatic theory', *Journal of Pragmatics,* 40: 1353-1356.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Sapir, E. (1971) [1921], *Language. An Introduction to the Study of Speech*. London: Rupert Hart-Davis.

Sinclair, J. (2004), *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stubbs, M. (2007a), 'Quantitative data on multi-word sequences in English: the case of the word *world*', in: M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert, *Text, Discourse and Corpora*. London: Continuum. 163-189.

Stubbs, M. (2007b), 'An example of frequent English phraseology: distributions, structures and functions', in: R. Facchinetti (ed.) *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi. 89-105.

Stubbs, M. (forthcoming), *Quantitative Data on Multi-word Sequences in English: The Case of Prepositional Phrases*. Lecture given at the Berlin-Brandenburgische Akademie der Wissenschaften, 3 November 2006.

Traugott, E.C. (1982), 'From propositional to textual and expressive meanings: some semantic-pragmatic aspects of grammaticalization', in: W.P. Lehmann and Y. Malkiel (eds.) *Perspectives on Historical Linguistics*. Amsterdam: John Benjamins. 245-271.

Traugott, E.C. and R.B. Dasher (2005), *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Whitsitt, S. (2005), 'A critique of the concept of semantic prosody', *International Journal of Corpus Linguistics,* 10(3): 283-305.

Wray, A. (2002), *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Research on fiction dialogue: Problems and possible solutions

Karin Axelsson

University of Gothenburg

Abstract

It seems quite clear that there must be differences in the language of fiction dialogue and the narrative parts of fiction, but fiction is often treated in corpus linguistics as if it were a homogenous genre, and there is very little quantitative research on the language of direct speech in modern fiction. The main problem is that corpora are seldom annotated for direct speech, and if they are, the mark-up may be difficult to use. If there is no mark-up for direct speech, corpus query matches have to be categorized and sorted manually as being inside or outside direct speech, and the proportion of direct speech in the corpus fiction texts needs to be investigated using, for example, statistical methods. As these procedures are time consuming, there is a need for specially designed corpora where direct speech is annotated. Another problem is how to define direct speech, slightly different definitions may be applied depending on whether comparisons are to be made to the narrative parts of fiction or to real-life speech. A further problem is that existing corpora are not usually sampled with a view to providing a representative sample of fiction dialogue; it seems important that samples are taken from different kinds of books and from different parts of books.

1. Background

In my overall research project, which is based on data from the *British National Corpus* (BNC), I study the use of tag questions in fiction dialogue and make comparisons with real-life speech. This has made me aware of some problems with existing corpora when they are used for research on fiction dialogue. The complication is that fiction texts are usually an integrated mix of two sub-genres: narrative and dialogue. Fiction dialogue here refers only to direct speech (the reporting clauses are not included). The term 'direct speech' will be used in this paper. Here is an example of fiction text, with direct speech in italics:

(1) She didn't look at her pursuer, who stopped inches away from Peter.
"*She's with you?*"
Peter tried to speak, failed and cleared his throat.
"*She certainly is*," he said in a voice that hardly quavered at all.
"*Okay. Just wondered. Always worth a try. You never know, do you?*"
"*No*," said Peter. "*You don't.*" (BNC: CKB 1365-1374)

Biber (1990) has studied inter-sample variation of some common grammatical features in different genres by comparing 1,000-word parts of the same samples in LOB (and the *London-Lund Corpus*):

> [M]ost of the difference scores are quite small across sub-samples […] conversation and general fiction show relatively large differences with respect to the pronominal, contracted, and tense features […] probably reflect[ing] changing purposes within the course of a text: for example […] shifts from narrative to description to dialogue within fiction. (Biber 1990: 259-261)

In spite of these "shifts", fiction is nevertheless often treated as if it were a homogeneous genre, although the direct speech parts and the narrative parts have very different purposes, as only the former parts try to mimic spoken language.

There seem to be rather few corpus-based studies on the language of direct speech in modern fiction. Oostdijk (1990) deals primarily with the ways in which direct speech is reported in fiction but she also enumerates some typical traits of direct speech in fiction without giving any quantitative data. She reports that "[t]ypical for direct speech […] was the use of *and*, *or*, and *but* sentence-initially as connective adjuncts" and also "the high [sic] frequent use of imperatives, interrogatives (especially tag questions) and exclamatory phrases, and of course such items as vocatives, interjections, clitic forms and responsive phrases", and she also found in direct speech various forms of ellipsis, e.g. omission of question operators in interrogative sentences, omission of subjects in declarative sentences, and unfinished sentences at the end of turns, as well as "numerous instances of incomplete sentences and loose phrases" (Oostdijk 1990: 239). Oostdijk also remarks that topicalization is frequent and that she found "creative use […] of substandard vocabulary and syntax to characterize the speech of some of the characters" (1990: 239-240). de Haan (1996) deals mainly with reporting verbs in fiction but he also presents data on the number of sentences with only direct speech, with only non-direct speech and with a mix of direct and non-direct speech, as well as the mean length of these types of sentences. de Haan (1997) finds that reported speech in fiction is characterized by a larger variation in sentence types than non-dialogue, as well as more extensive use of the present tense and more occurrences of marked word order (especially preposed subject complements and preposed direct objects). However, it should be noted that de Haan's results are based on data from a very limited number of novels: de Haan (1996), seven samples, and de Haan (1997), three samples, in the latter case all of them crime novels.

It is remarkable that so little quantitative research on the language of direct speech in fiction has been published so far, as it is a type of language we all encounter quite often. Direct speech in fiction is said to give "the illusion of real speech" (Leech and Short 2007: 132), which makes it interesting to find out in what ways and how much the language of direct speech in fiction differs from real-life speech, apart from the obvious absence in fictional direct speech of

"features of normal non-fluency" often found in real-life speech, such as hesitation pauses, false starts and syntactic anomalies (Leech and Short 2007: 130). General studies of direct speech in fiction may also provide reference material for the study of direct speech in certain literary genres, for certain authors, or even for certain individual works, as "'creativity' can only be recognized as such when there is a language norm against which the 'creative' language comes to stand out" (Mahlberg 2007: 221).

More research on direct speech in fiction as well as research on the narrative parts of fiction would give a fuller picture of 'the language of fiction'.

2. The problems

The main problem when performing research on the language of direct speech in fiction is that there are very few corpora where direct speech is annotated as such. Searches cannot usually be restricted to direct speech and there is no information on the number of words in direct speech, so frequency calculations are difficult to perform. The main reason for the dearth of research on direct speech in modern fiction is probably this lack of specially adapted corpora. Most English corpora with samples from fiction, e.g. the BNC, are not tagged for direct speech.

Another problem is how to define direct speech. If one wants to compare direct speech in fiction with real-life speech, direct thought should not be included. A definition of direct speech for this purpose could then be: direct speech presents verbatim what a character is claimed to have said in the fictional world (cf. Semino and Short 2004: 12). This means that hypothetical direct speech should also be excluded, as in the italicized part of example (2):

(2) A love song […] will make tears come to my eyes: move me with the desire to say, *You do love me, don't you! This love will last for ever. This love, lasting forever, makes me immortal. This love replaces death*. I don't say it of course. (BNC: HGJ 435-438)

If the purpose, on the other hand, is to compare the use of direct speech to the narrative parts in the fiction texts, hypothetical speech should probably be included, and direct thought and reporting clauses might have to be annotated as such to enable searches restricted to the purely narrative parts of the texts.

Another issue is the representativeness of the fiction samples. One cannot take it for granted, for example, that the amount of direct speech or the linguistic features of direct speech are similar in samples from different parts of fiction texts (beginning, middle and end samples). There may also be differences in direct speech between different literary genres (see e.g. de Haan 1996 on a very high amount of direct speech in some crime novel samples). Moreover, the use of direct speech might also vary between novels depending on their literary quality. The problem here is that literary quality is difficult to measure objectively, but attempts have been made to categorize fiction texts as having different levels of

perceived difficulty (Burnard 2000) and to make a distinction between serious and popular fiction (Semino and Short 2004: 21).

There are several solutions which more or less solve these problems. First, one may use existing corpora with some mark-up of direct speech. Second, methods may be applied which makes it possible to use corpora without any mark-up of direct speech. Third, already existing corpora may be annotated for direct speech, and finally, new specially designed corpora may be created. These four solutions will be discussed in the following sections.

3. The first solution: using existing corpora with some mark-up of direct speech

There are, in fact, some corpora with mark-up of direct speech, which shows that annotation of direct speech is quite feasible. However, this does not mean that there are no problems.

A corpus where direct speech has been tagged is the *English-Norwegian Parallel Corpus* (ENPC): in the original fiction parts of ENPC (ca 800,000 words), direct speech (including direct thought) has been annotated, partly automatically (using quotation marks) and partly manually (Johansson et al. 1999/2002). It is very easy to restrict searches to direct speech in the ENPC as there is a tickable box for this on the search page. However, there are no figures for the total number of words in direct speech in the ENPC, so frequency calculations are still problematic.

The corpora in the *Brown Corpus* family have some mark-up of metatextual information including indications of direct speech quotations but Leech and Smith (2005) report that, according to their experience,

> few users of these corpora manage to exploit the potential of the mark-up […]. The problem partly stems from differences of mark-up conventions from one corpus to another […]. In addition, there has been a shortage of generally available software to allow easy exploitation of these mark-up codes. (Leech and Smith 2005: 92)

This mark-up of direct speech is not mentioned in the *Brown Corpus Manual* (Francis and Kucera 1979). Furthermore, the information in the manual shows that the fiction samples in the *Brown Corpus* were selected randomly from different parts of novels but "no samples were admitted which consisted of more than 50% dialogue" (Francis and Kucera 1979), which of course affects the representativeness of the fiction samples in the *Brown Corpus* (and probably also in the other corpora of the family as the sampling criteria are supposed to be the same for all corpora in the *Brown Corpus* family).

There are also other corpora where direct speech has been annotated, especially historical corpora, since there is a special interest within historical linguistics in written representations of spoken English as this is the only way of

getting data on speech-like language from historical periods. Examples of such annotated fiction texts can be found in e.g. *A Corpus of English Dialogues* (http://www.engelska.uu.se/corpus.html, accessed 7 July 2008) and the ARCHER corpus (where the annotation was made manually, Douglas Biber, p.c.).

4. The second solution: using existing corpora without any mark-up of direct speech

Although there are some corpora with mark-up of direct speech, this is generally not the case. This holds for the BNC, which I use in my research, and apparently also for, among others, the ICE corpora (http://www.ucl.ac.uk/english-usage/ice/annotate.htm, accessed 7 July 2008) and *the Bank of English* (http://www.titania.bham.ac.uk/docs/svenguide.html, accessed 7 July 2008).

If one wants to do research on direct speech in fiction using existing corpora without annotation of direct speech, one needs to manually discard all matches outside direct speech, and in order to study frequencies, calculate the proportion of direct speech.

There are some studies on the proportion of direct speech in fiction. de Haan (1996) reports around 40 percent direct speech in a 140,000-word corpus (seven samples of 20,000 words each). Semino and Short (2004: 67-68) present figures for speech presentation in a specially designed corpus, the *Speech, Writing and Thought Presentation Written Corpus*, and from these figures one may calculate the proportion of direct speech in their fiction material (87,709 words) at between 22 and 23 percent.

The proportions of direct speech are thus very different in these two studies. This is probably due to the relatively small size of the corpora and how the samples were selected. The proportion of direct speech seems to vary widely between different works of fiction (see e.g. de Haan 1996: 27). In order to find out the proportion of direct speech in the corpus or subcorpus used in a study of direct speech in fiction, one may make a statistical investigation of the proportion of direct speech in that particular material by classifying randomly selected words as being inside or outside direct speech. Of course, such statistical figures will have a margin of error, but they may still be the most reliable figures one can obtain. A more detailed description of how such a statistical investigation of direct speech can be performed is presented in the next section.

4.1 A statistical investigation of direct speech in fiction

For my study on the use of tag questions in direct speech in fiction, I use a BNC subcorpus of about 9.7 million words which is restricted to Lee's genre fiction prose, book as medium, UK and Ireland as domicile of author and a publication date of between 1985 and 1993. In order to make a statistical investigation of the proportion of direct speech, one needs to have access to the full text files in the corpus. I therefore used the XML-files found on the DVD to the BNC XML Edition.

The procedure was as follows. All w-units in the 262 files in the subcorpus were counted and given an individual number (multiword-units were considered as w-units, as they were tagged as such in the BNC World version used in the rest of the project). Then a program with a random generator selected 2,000 of the 9,711,449 numbers, and extracted the s-units with the w-units with those numbers from the 262 files together with five s-units of context before and five s-units of context after each of these s-units. The 2,000 extracts were then exported to pdf-format, from where they were then printed out for ease of analysis. The randomly selected w-units were marked in bold in the extracts. These 2,000 w-units were then classified as being within or outside direct speech using the definition given in section 2. In some cases, it was necessary to check a larger context in the corpus to decide if the selected w-unit was found in direct speech or not.

638 out of the 2,000 w-units were found to be within direct speech, which gives a proportion of 31.90 percent (margin of error: ± 0.204 percent units at $p < 0.05$). This proportion was then used to calculate how many of the w-units in the whole subcorpus are inside direct speech, which enabled me to compute the frequency of tag questions in direct speech in my BNC Fiction Subcorpus and compare it with the frequency of tag questions in the spoken demographic part of the BNC (Axelsson forthcoming).

5. The third solution: annotating direct speech in existing corpora

Another solution would be to annotate direct speech in already existing corpora. This is a larger project than making a statistical investigation and often practically impossible for the individual corpus user. If such improvements were added directly in already existing corpora, or if tools for such annotation which were easy to apply for the individual corpus user were available, research on direct speech would be greatly facilitated.

One remaining problem, however, is that fiction corpora (or fiction components of corpora) may not be balanced to give a representative picture of direct speech in fiction.

As mentioned above, the compilers of the *Brown Corpus* selected samples randomly from different parts of novels: "The page on which to begin the sample was also selected by the random number table" (Francis and Kucera 1979: section 1). One would suppose that the *Brown Corpus* should have set a standard here, but later corpora often seem to select samples less randomly. One common principle appears to be taking only beginning samples. This is the case in e.g. the *English-Norwegian Parallel Corpus* (Johansson et al. 1999/2002), where the first 10,000-15,000 words are included, and the *Corpus of Contemporary American English* (COCA), where only first chapters are used (Davies 2009).

In the *British National Corpus*, samples are claimed to be "taken randomly from the beginning, middle or end of longer texts" (Burnard 2000), but figure 1 shows the actual distribution (in w-units) in my BNC Fiction Subcorpus. Middle

samples dominate, which seems natural, but there is an imbalance between beginning and end samples.

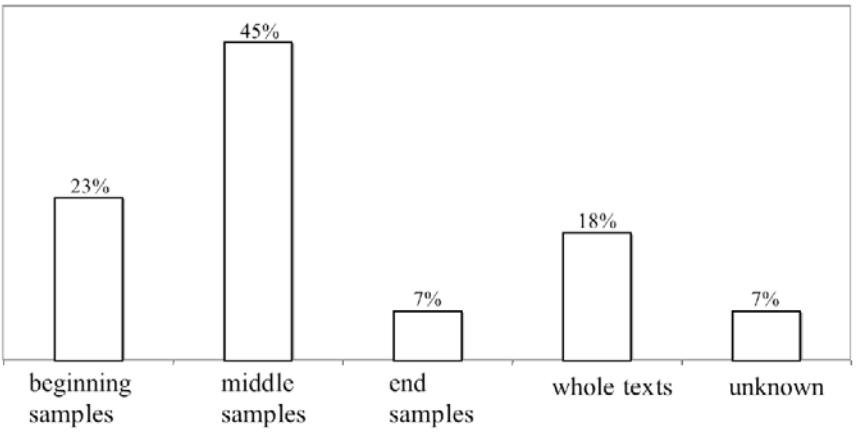

Figure 1: Percentage of w-units in different sample types in my BNC Fiction Subcorpus

This imbalance might have been due to the search restrictions applied for my subcorpus, but this imbalance is even more pronounced if we consider all fiction prose book samples in the BNC, as shown in figure 2.

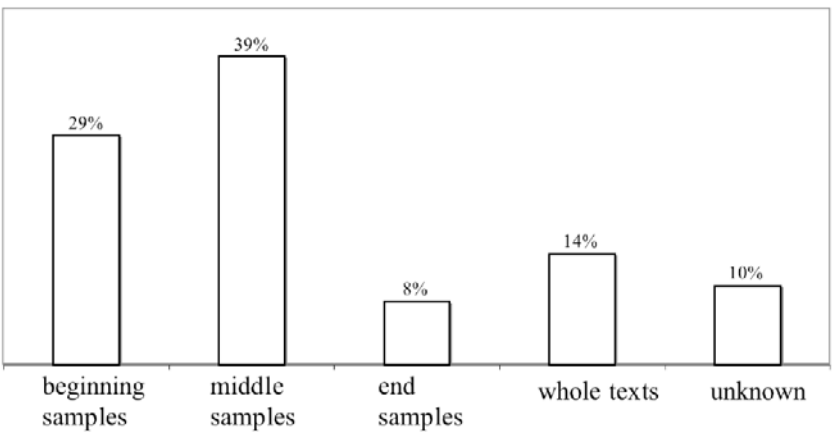

Figure 2: Percentage of w-units in different sample types in all fiction prose in the BNC

My statistical investigation of direct speech shows no significant differences in the proportion of direct speech in different sample types, but this may be due to the small size of the statistical sample. However, there seem to be more tag questions in end samples than in beginning and middle samples, but the number of end samples is quite small and there are indications that another factor may also be involved, namely the 'level of difficulty' (Burnard 2000).

Perceived level of difficulty is described in the BNC manual as "a subjective assessment of the text's technicality or difficulty" which formed part of an attempt "to characterize the kind of audience for which written texts were produced" (Burnard 2000); Burnard admits that the level of difficulty "proved very difficult to assess and was very frequently confused with circulation size or audience size". Level of difficulty is a descriptive feature in the BNC, not a selection feature, which means that it was not used as a criterion when including samples in the corpus. The selection of samples from books in the BNC was partly random and partly based on bestseller shortlists and library statistics. There are three levels of perceived difficulty in the BNC: high, medium and low. Figure 3 shows the distribution of these levels in my BNC Fiction Subcorpus.

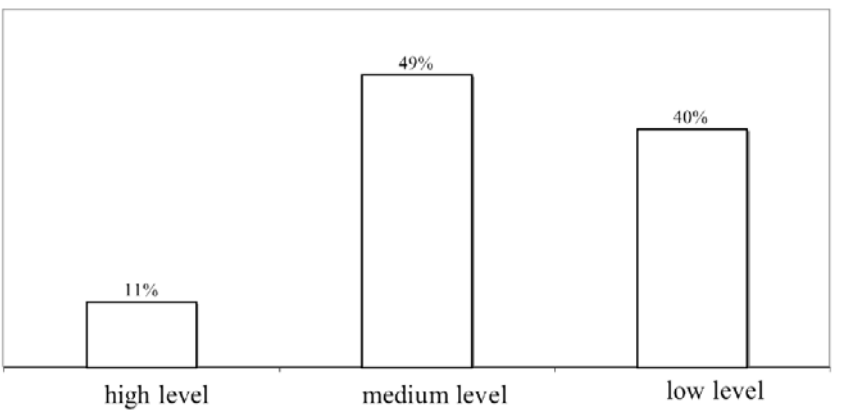

Figure 3: Percentage of w-units in samples of different levels of difficulty in my BNC Fiction Subcorpus

The distribution of samples from different parts of the books is far from random within these categories – the imbalance is greatest for samples from novels with a high level of difficulty: there is only one end sample in this category. The samples from novels with a low level of difficulty, on the other hand, display a balanced distribution from different parts of the books.

Another qualitative categorization of fiction samples has been made in the *Speech, Writing and Thought Presentation Written Corpus,* where half of the samples are from 'serious fiction' and the other half from 'popular fiction', as the corpus team wanted to compare the speech, writing and thought presentation in these types of books (Semino and Short 2004). Semino and Short realized that this classification could be controversial but "felt that ignoring it could have re-

sulted in a biased or unbalanced choice of texts" (2004: 22). Their classification is based on a combination of the classifications in the BNC and the Oxford Text Archive, the personal opinions of nine members of Lancaster University's Stylistics Research Group, an interpretation of publishing and marketing strategies and whether the novels have been shortlisted for literary prizes (Semino and Short 2004: 23).

Some books and authors are represented in both the BNC and the *Speech, Writing and Thought Presentation Written Corpus*, which makes it possible to compare the categories. BNC's 'high' seems to correspond roughly to Semino and Short's 'serious fiction', and BNC's 'medium' to Semino and Short's 'popular fiction', whereas BNC's 'low' appears to be missing in Semino and Short's corpus.

The statistical investigation of direct speech in my BNC Fiction Subcorpus shows that there is a significantly lower proportion of direct speech in samples with a high level of perceived difficulty. Figure 4 shows this distribution of matches of direct speech per million w-units in the statistical sample.

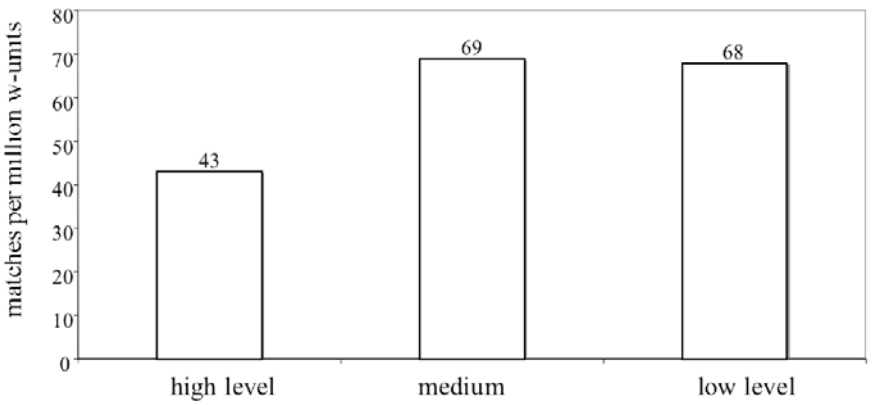

Figure 4: Direct speech matches per million words in the statistical investigation of the proportion of direct speech in samples of different levels of perceived difficulty

The difference between serious and popular fiction is even larger for the *Speech, Writing and Thought Presentation Written Corpus*. Figures in Semino and Short (2004: 67-68) make it possible to calculate the proportion of direct speech (in words) in popular fiction at 29-30 percent, whereas serious fiction contains only 15-16 percent. The difference in the proportion of direct speech between the two corpora (see section 4) is probably partly due to different balancing of samples from different types of literature. My research on tag questions in fiction indicates that tag questions are less frequent in samples with a high level of perceived difficulty: this is probably partly due to there being less direct speech in such books (Axelsson forthcoming).

6. The fourth solution: creating new specially designed corpora

A corpus specially designed for research on direct speech in fiction should preferably:

- be tagged for direct speech, so that searches can be restricted to direct speech;
- give the total number of words in direct speech;
- provide a balanced random sampling from different parts of the fiction texts – preferably beginning, middle and end samples from all included texts;
- be balanced for different levels of difficulty (probably quite difficult, but obviously necessary);
- be balanced for literary genres, and allow comparisons of direct speech between these genres.

If corpora are designed specially for research on direct speech in fiction, the question is whether some kinds of books should be excluded. It is problematic if novels with a historical setting are included in a corpus of contemporary fiction as the language in direct speech in such novels is sometimes archaized, as in examples (3) and (4):

(3) "Methought you knew all about it," pointed out Joan. (BNC: CCD 339) (from *The Child Bride* by Philippa Wiat, set at the time of the War of the Roses (http://www.r3.org/fiction/roses/murph.html, accessed 8 July 2008))

(4) "Ask me no more, Rose, I pray you!" (BNC: HGV 4349) (from *Hidden Flame* by Elizabeth Bailey, set in 1797 (http://historicalromancewriters.com/Bookinfo.cfm?bookID=22497, accessed 8 July 2008))

Tag questions are an illustrative case of this. Question tags usually have enclitic negation; non-enclitic negation in question tags is mostly found in formal English or northern British dialects (Quirk et al. 1985: 810). Question tags with non-enclitic negation are thus quite rare in Present-day spoken English, and they prove to be quite rare also in modern fiction: only 20 out of 649 negative question tags found after declarative anchors in a thinned version of my BNC Fiction Subcorpus have non-enclitic negation. However, at least half of these are from novels with a clear historical setting, as in examples (5) and (6). The inclusion of such samples from historical novels thus affects the corpus results.

(5) "My prince meanwhile is heir presumptive, is he not?"
(BNC: CCD 995) (from *The Child Bride* (see example (3))

(6) "Here we are to spend a merry night, are we not, Thomas?"
(BNC: H8A 330) (from *Murder Makes an Entrée* by Amy Myers, set during the Edwardian era (http://www.fantasticfiction.co.uk/m/amy-myers/murder-makes-entree.htm, accessed 8 July 2008))

Another question is whether samples from novels which break the norms of fiction should be excluded from a corpus of fiction. Example (7) is an extract from *Talking It Over* by Julian Barnes.

(7) "I'm not very good at telling."
There was a half-eaten chicken tikka in front of her and a half-drunk glass of white wine. Between us stood a fat red candle, whose flame was beginning to drown in a pond of wax, and a purple African violet made of plastic. By the light of that candle I looked at Gillian's face, properly, for the first time. She … *well, you've seen her for yourself, haven't you? Did you spot that tiny patch of freckles on her left cheek? You did?* Anyway, that evening her hair was swept up over her ears at the sides and fastened back with tortoise-shell clips, her eyes seemed dark as dark, and I just couldn't get over her. I looked and I looked as the candle fought with the wax and cast a flickering light on her face, and I just couldn't get over her.
"I don't either," I finally said. (BNC: EDJ 711-720)

In the middle of the narrative, the narrator suddenly seems to address the reader with questions and it is even hinted that the questions are answered before the narrator goes back to telling the story in a more normal way. The part in italics is a kind of direct speech, but it is not something which is uttered in the fictional world. Such direct speech is excluded in my study of tag questions. When creating a corpus of fiction, it is important to decide if the corpus is meant to reflect typical fiction or also a range of books which in different ways break the norms of fiction and thus constitute atypical fiction.

7. Summary and conclusion

This paper has pointed at some problems when performing corpus research on fiction dialogue, the main one being that direct speech is usually not tagged as such in corpora of modern English. The best solution would be to create specially designed corpora where the issues of representativeness could also be addressed. While waiting for such corpora to be created, already existing corpora could be annotated for direct speech manually. Another solution described in this paper is to discard all examples found outside direct speech and then perform a statistical investigation of the proportion of direct speech in the corpus or subcorpus used.

Linguistic research on direct speech in fiction is greatly facilitated if it is possible to restrict corpus searches to direct speech and if the size of the direct speech parts of the corpus (or subcorpus) is known. Results from corpus investigations of direct speech in fiction would probably be more reliable if efforts were made to balance such corpora for different literary genres, different levels of literary quality and for different parts of novels.

Acknowledgments

I would like to thank all those who helped me with the statistical investigation of direct speech: linguists at ICAME 28 in Stratford-on-Avon 2007, especially Stefan Evert, for giving me advice on the principles for a statistical investigation of direct speech, and, in particular, Robert Andersson, system administrator and IT-coordinator at the Swedish National Graduate School of Language Technology at the University of Gothenburg, for helping me with the programming. I am also grateful to The Royal Society of Arts and Sciences in Göteborg for a grant enabling me to take part in ICAME 29 in order to present a poster on this topic.

Corpora

British National Corpus, version 2 (BNC World) (2001), Distributed by Oxford University Computing Services on behalf of the BNC Consortium, http://www.natcorp.ox.ac.uk.

British National Corpus, version 3 (BNC XML Edition) (2007), Distributed by Oxford University Computing Services on behalf of the BNC Consortium, http://www.natcorp.ox.ac.uk.

Davies, M. (2009), *Corpus of Contemporary American English (COCA)*, http://www.americancorpus.org, accessed 5 February 2009.

References

Axelsson, K. (forthcoming), *A Corpus-based Study of the Use and Function of Tag Questions in Direct Speech in Fiction Compared to Real-life Speech*. PhD thesis, University of Gothenburg.

Biber, D. (1990), 'Methodological issues regarding corpus-based analyses of linguistic variation', *Literary and Linguistic Computing*, 5(4): 257-269.

Burnard, L. (2000), *Reference Guide for the British National Corpus (World Edition)*, http://www.natcorp.ox.ac.uk/docs/userManual, accessed 31 March 2008.

de Haan, P. (1996), 'More on the language of dialogue in fiction', *ICAME Journal*, 20: 23-40.

de Haan, P. (1997), 'Syntactic characteristics of dialogue and non-dialogue sentences in fiction writing', in: M. Ljung (ed.) *Corpus-based Studies in English: Papers from the 17th International Conference on English Language Research on Computerized Corpora (ICAME 17)*, Stockholm, May 15-19, 1996. Amsterdam: Rodopi. 101-117.

Francis, W.N. and H. Kucera (1979), *Brown Corpus Manual*, http://khnt.aksis.uib.no/icame/manuals/brown, accessed 2 May 2008.

Johansson, S., J. Ebeling and S. Oksefjell (1999/2002), *English-Norwegian Parallel Corpus: Manual,* http://www.hf.uio.no/ilos/forskning/forsknings prosjekter/enpc, accessed 2 May 2008.

Leech, G. and M. Short (2007), *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2nd edition. Harlow: Pearson.

Leech, G. and N. Smith (2005), 'Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB', *ICAME Journal*, 29: 83-98.

Mahlberg, M. (2007), 'Corpus stylistics: bridging the gap between linguistic and literary studies', in: M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, Discourse and Corpora*. London: Continuum. 219-246.

Oostdijk, N. (1990), 'The language of dialogue in fiction', *Literary and Linguistic Computing*, 5(3): 235-241.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language.* Harlow: Longman.

Semino, E. and M. Short (2004), *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Establishing the EU: The representation of Europe in the press in 1993 and 2005

Anna Marchi and Charlotte Taylor

Universities of Lancaster and Siena

Abstract

This paper investigates how the European Union was represented in three British newspapers over two different time periods: 1993 and 2005. The Treaty on European Union*, which led to the creation of the European Union, was signed in 1992 and entered into force in 1993.* The Treaty Establishing a Constitution for Europe *was signed in 2004, and like the* Maastricht Treaty *was subject to ratification. However, unlike the* Maastricht Treaty*, it was rejected in referendums in France and the Netherlands in 2005 and therefore was not implemented. These two events were chosen for their importance in the history of the European Union and because they allow for a diachronic comparison of the construal of Europe in the British press. Two sub-corpora were used in the study, the first,* SiBol_93*, contains approximately 92 million tokens from three broadsheet British newspapers collected in 1993 and the second,* SiBol_05*, contains approximately 150 million tokens collected from the same sources in 2005. Each of these corpora covers the year after the signing of the treaties and therefore the period in which the ratification was discussed. The corpora were investigated using Corpus-Assisted Discourse Studies (CADS) which involves a shunting between quantitative and qualitative analytical approaches and starting points (see, for example, Partington 2004, forthcoming; Baker 2006). Our findings show that while there is no simplistic positive to negative reversal of evaluation, there is certainly a marked decrease in the newsworthiness of Europe and the European Union, and the problem the European Union faces is primarily one of visibility.*

1. Introduction

1993 and 2005 were key years for the European Union: in 1993 the *Treaty on European Union* (commonly referred to as the *Maastricht Treaty*) which established the EU was ratified; while in 2005 the *Treaty Establishing a Constitution for Europe* (commonly referred to as the *European Constitution*) also went through the ratification process. While the 1993 treaty was successfully passed by all the member states after a difficult ratification stage, the 2005 treaty was abandoned after being rejected in referendums in both France and the Netherlands. A brief timeline has been included for ease of reference (figure 1).

Despite the obvious difference in outcome, there were similarities in the treatment of the two treaties as in both cases there was strong opposition in the British press. By focussing on the two treaties we hoped to bring out some of the linguistic means used to construe Europe in the two time periods, and also to see how the European Union was established. The diachronic aspect is particularly interesting because there has been such an enormous shift in attitudes towards European union and the European Union in the British press, from the much quoted *Sun* editorial in 1975, which, in the referendum on membership of the European Economic Community, told its readers "You can vote YES – FOR A FUTURE TOGETHER Or NO – FOR A FUTURE ALONE", to the present day situation.[1]

1973	Britain, Ireland and Denmark join the EEC
1990	UK enters the ERM (Exchange Rate Mechanism)
1991	Treaty on European Union (Maastricht Treaty) signed (subject to ratification by member states).
	Treaty on European Union rejected in a referendum in Denmark
	Treaty on European Union narrowly accepted in a referendum in France
1992	British Pound, Italian Lira and Spanish Peseta forced to exit the ERM
1993	Treaty on European Union ratified by Denmark
	Treaty on European Union enters into force
1999	Start of the third stage of European Monetary Union: 11 EU countries adopt the euro
2002	Euro notes and coins are introduced in 12 EU countries.
2004	Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia join the European Union
	Treaty Establishing a Constitution for Europe (TECE) signed in Rome (subject to ratification by member states).
2005	TECE rejected in a referendum in France
	TECE rejected in a referendum in the Netherlands

Figure 1: Timeline of key events in the creation of the European Union

Our aim was to create a background and to experiment with ways of addressing the question of Europe in the press, which would help us in our work on more recent media attitudes towards the European Union and European identity as part of a much larger research project: *IntUne*.[2] Like recent work described in Millar and Wilson (2007), we take the view that "Europe is not a pre-determined entity waiting to be discovered, but something that is socially and discursively 'constructed'" (2007: 7). Previous research in the area has frequently focussed either on a quantitative investigation, such as Peter and de Vreese's (2004) extensive study into TV news across Europe, or a qualitative investigation, such as Firmstone's (2003) detailed survey of UK editorial coverage of the introduction of the Euro, while our intention was to combine the two approaches, and to primarily concentrate on the ways in which Europe is talked about. Where

possible, we have included several illustrative and extended examples to allow the reader to check our interpretations, and to draw his/her own conclusions.

2. Methodology

This study finds its methodological and theoretical base in the relatively recent tradition of Corpus-Assisted Discourse Studies (CADS). CADS aims at identifying distinctive features and investigating non-obvious meanings within specific discourse types. The analysis is in its nature comparative, our interest here is mainly diachronic and this work is based on the comparison of two specialised sub-corpora of modern newspaper texts, compiled at twelve years' distance.

In Modern Diachronic CADS, qualitative and quantitative methods are combined in order to analyse comparable corpora from different (although recent) times (Partington and Duguid 2008). This kind of investigation allows us both to track changes in language usage and to account for extralinguistic (social, political, historical, cultural, etc.) changes that language reflects. CADS inherits from corpus linguistics the inductive approach, a bottom-up, data-driven analysis that relies on a large amount of linguistic data. But CADS also draws from the discourse analysis approach, stressing the importance of in-depth analysis of texts, as well as the importance of looking at the context outside the concordance lines, using a variety of starting points and inputs, i.e. looking 'inside the box' rather than adopting a *faux naïf* stance.

There is a growing literature praising the synergic combination of qualitative and quantitative methods in applied linguistics (Hardt-Mautner 1995; Stubbs 2001; Partington 2004; Baker et al. 2008) and an increasing number of studies adopt corpus linguistic techniques in order to improve text analysis or, conversely, find explanations for the patterns emerging from the corpus in the rich contextual and theoretical framework of qualitative approaches. Partington writes:

> Complementing the qualitative with a more quantitative approach, as embodied in Corpus Linguistics, not only allows a greater distance to be preserved between observer and the data but also enables a far greater amount of data to be contemplated. In addition, it can identify promising areas for qualitative forms of analysis to investigate. (Partington 2006: 268)

Quantitative and qualitative approaches interact and add to each other throughout the analysis, the process is recursive: we start with a research question, we turn to the data (starting from wordlists and keywords) and look for patterns, when we find something potentially interesting we look more closely, expanding the analysis to the larger co-text (enlarged concordance lines or whole texts) and, taking into account contextual elements, close reading and extralinguistic

elements implement the original research question and open new questions, and the process goes on as the plot thickens.

3. The corpus

For this study we used the *SiBol Corpus*, the *Siena-Bologna Modern Diachronic Corpus* of British newspapers. The corpus is composed of two sub-corpora, containing the whole output of the main British national quality papers, in 1993 and in 2005 respectively. The composition of *SiBol_93* and *SiBol_05* is perfectly symmetric, we have the complete final edition of the same newspapers (*Guardian*, *Daily Telegraph*, *Sunday Telegraph*, *Times* and *Sunday Times*) for the two years under investigation. *SiBol_93* contains 98,943,032 orthographic words, *SiBol_05* is bigger, with 146,077,408 orthographic words.

The corpus is XML[3]-valid and TEI[4]-conformant, all texts have been annotated in order to retrieve specific news types or portions of the newspapers (e.g. news pages vs. comment pages or front pages vs. supplements) and specific portions of individual articles (e.g. headlines or first paragraphs). Mark-up gives us access to some basic contextual information about the newspapers and about the articles (e.g. publication, section, by-line) and allows us to interrogate the corpus with a greater level of delicacy. The newspapers have been encoded according to a variety of parameters, this categorisation allows us to make comparisons across the defined partitions, we thus can easily compare the newspapers at a diachronic level, or on the basis of their political orientation, or at the specific level of each individual newspaper.

The XML corpus has been indexed and interrogated using Xaira[5] software, which has been used for collocation and concordance analysis. Wordlists and keyword lists have been produced using WordSmith 5.0.[6]

4. Analysing Europe: starting points

4.1 Editorials: reading the 'voice of the newspaper'

As noted in the methodology section above, CADS involves looking at 'non-obvious meaning', but to do so we need to pay attention to the 'obvious meaning', to avoid the trap of reinventing the wheel. Therefore, in order to explore what each of the newspapers was explicitly saying about Europe, another starting point was formed by simply reading a sample of the editorials from the two corpora: *SiBol_93* and *SiBol_05*. In each case, the headline and first paragraph were scanned for words associated with Europe and approximately 25 articles for each year which were considered (subjectively) to be 'about' Europe were then read in full.

Editorials were chosen as these represent the 'voice of the newspaper', and as Morley (2004: 239) summarises "[t]hey classically comment on recent events

which are narrated elsewhere in the newspaper's news stories. One of the prime functions of editorial comment is that of persuading the newspaper's readers of its point of view". Therefore, they seemed an ideal way of focussing on overt and obvious meanings. Clearly, that is not to imply that this is the only voice in the newspaper, as reporters and columnists may dissent from the official position, as Anderson (2004: 152) notes in his discussion of English press attitudes: "[t]he Eurosceptic press consists of those papers in which Euroscepticism is the dominant, but not necessarily the exclusive discourse on the EU", and this also applies, vice versa, to pro-European or Euro-friendly newspapers.

Reading the editorials showed that, in 1993, the terms *EU*, *EC*, and *EEC* were certainly not treated as interchangeable in the three newspapers, and in their editorials the newspapers are explicit about the choices made in "naming Europe", as the *Times* describes it in the first example below: [7]

(1) Naming *Europe*
Who will call the Twelve a *union*? On November 1 a new political creature was born in the heart of Europe and no one yet knows what to call the mewling infant. Its *Eurocrat* godparents have christened it the *European Union*; but others think that name more appropriate for an insurance company, a continental motorway junction or an international gathering of Christian mothers. (Editorial, *Times*, 13/11/1993)

(2) Black Monday
THIS is the last time you will read about the *European Economic Community* in The Sunday Telegraph. From tomorrow – All Saints' Day – the EEC is officially abolished. Before you rush out to celebrate, however, you should be aware that the EEC is not dead. It has simply changed its name, under the Maastricht Treaty, to something even more grandiose. It is now formally called the *European Community*. Since this is now the legal position, The Sunday Telegraph will follow that style in future.
Something else comes into being tomorrow. It is the *European Union*, also created by Maastricht. (Editorial, *Sunday Telegraph*, 31/10/1993)

(3) Paying the Union dues
THOUGH you may not have noticed, Britain has for over a week now been part of the *European Union*, of which we are all automatically citizens: even the Queen […]. The Guardian, while not putting out flags, intends to honour the logic of the occasion and use the expression *European Union*, and the short form *EU*, where it correctly applies. We do this for two reasons. First, on the good old principle that a spade should be called a spade and that people and institutions addressed by the names they use themselves […]. Second, because as it happens *EC* occupies a particularly cluttered sector of the acronymic forest. (Editorial, *Guardian*, 9/11/1993)

In addition to highlighting the open discussion regarding the choice of terms, these examples also serve to show the overt evaluation of the newly formed European Union in 1993, which was somewhat less than enthusiastic in all three

newspapers. In the examples above, it is described as the "mewling infant" of "Eurocrat godparents" in the *Times*, while the *Telegraph* assumes that its readers would celebrate its demise, and even the *Guardian* struggles to come up with convincing or ideological reasons for adopting the new name, po-facedly noting that "as it happens *EC* occupies a particularly cluttered sector of the acronymic forest".

Overt evaluation was also frequent in the editorials in *SiBol_05*. Starting with the *Telegraph*, opposition to the European Union is very clearly stated, in one editorial the rejection of the *Treaty Establishing a Constitution for Europe* is described in very positive terms (example (4)):

(4) *The chance for a dynamic new start in Europe*
Today's meeting of European leaders is potentially the most momentous summit in the history of the European Union, if not of its predecessors, the EC and the EEC. We are witnessing the natural demise of the process of "ever closer union" that began in 1957 with the Treaty of Rome. The attempt to weld the ancient nations of Europe into one country, with all the "competencies" and characteristics of statehood, has run its course. *The opportunity before us, therefore, is immense.* (*Telegraph*, 16/06/2005)

Another editorial proposes a fairly lengthy mock constitution, described as a "Healthy Constitution" in the headline, which will reassert "the supremacy of national parliamentary and legal systems", thus summarising the *Telegraph*'s attitudes towards the European Union. In example (4) the *Telegraph* talks of the "natural demise" of the union, and in other editorials both the EU (example (6)), and more unusually, Europeans (example (5)) are described with metaphors of mental illness:

(5) As she has found before, and will find again, Europeans often exhibit a *psychotic* desire to bite the hand that freed them. (*Telegraph*, 05/12/2005)

(6) EU *schizophrenia*
Tony Blair was in Hampton Court this week talking, as usual, about economic reform. Listening to him, you might easily have got the idea that the EU was streamlining and deregulating. So it comes as quite a shock to look at the budget that Brussels has just passed. For all the Prime Minister's high-tech rhetoric, the EU plans to fire-hose money at the most corrupt, expensive, wasteful and immoral system of farm support on Earth. (*Telegraph*, 29/10/2005)

Example (6) also shows the *Telegraph*'s fierce opposition to the Common Agricultural Policy (CAP). The CAP was also opposed by the *Times*:

(7) Both sets of talks are deadlocked because of the CAP, whose subsidies *keep millions of farmers in the developing world mired in poverty*. It forms

> the core of an eccentric EU budget that robs the continent of diplomatic dignity and economic sanity. (*Times*, 14/12/2005)

In contrast with the *Telegraph*, the *Times* example in (7) above expresses the problem in very human terms, the focus on the "farmers" who are affected rather than the "system" itself. This tendency to focus on the human participants was seen in other editorials, see for example "voters" and "people" in (8), and suggests an interesting area for later research.

(8) The drubbing that French and Dutch *voters* gave the EU's needlessly complex, guilefully ambiguous, constitutional treaty — a document that, it must be remembered, Mr Blair vigorously endorsed and has yet to say that he no longer supports — may finally be bringing home to Europe's complacent *eurocrats* and politicians that *people* do not see why they should give *Brussels* ever more power, for purposes that they cannot entirely understand. (*Times*, 01/07/2005)

Example (8) additionally serves to illustrate the *Times*' heavily negative criticism of the *Treaty Establishing a Constitution for Europe* and the metonymic *Brussels*. However, the *Times* editorials were not purely negative about the European Union, for example in talking of Turkey (whose treatment is heavily criticised in both the *Times* and the *Guardian* editorials) in example (9):

(9) The EU's *great success* has been to provide economic and social benchmarks that must be met by new members, even if "old" Europe's economic vitality is itself in question. (*Times*, 01/07/2005)

In contrast to the other two broadsheets, the *Guardian* editorials show overt support for the *Constitution for Europe*, as in example (10):

(10) If we were a French newspaper we would be urging our readers to say yes. (*Guardian*, 28/05/2005)

However, there is simultaneously an accrual of dismissal of the European Union and its institutions, as illustrated in examples (11) and (12):

(11) It would be an exaggeration to say that millions of Europeans are waiting with bated breath to see how Austria handles the business of the European Union when it takes over the presidency tomorrow […]. Unsurprisingly, ambitions for 2006 are modest. No epoch-changing decisions are looming: the euro is up and limping and the "big bang" enlargement of May 2004 is reflected in French malaise, Polish builders in London suburbs and German anxiety about losing jobs to tax-cutting easterners. (*Guardian*, 31/12/2005)

(12) Messy reality and paltry results in Brussels are not inspiring – but governments still have a duty to keep on selling the EU's benefits to their doubting and apathetic people. (*Guardian*, 24/03/2005)

Having examined the editorials, the differences between the overt stance adopted by the three newspapers, and the attention that they paid to issues such as the naming of the union was clear, thus paving the way to look at more 'non-obvious' meanings, those that might not have been immediately available to the casual reader of the newspapers.

4.2 Keywords relating to Europe

The next starting point, and a very common one in the WordSmith (driven) world of corpus linguistics, was the calculation of keywords. Keywords are those items which are found to have a significantly high frequency in one corpus compared to another. The keywords shown in table 1 were calculated using WordSmith 5.0 (Scott 2008) to compare *SiBol_93* to *SiBol_05* and vice versa. Therefore, rather than using an 'external' reference corpus, such as the *British National Corpus*, which would have introduced unwanted variables, each corpus functioned as the other's reference corpus.

Table 1: Selected keywords concerning Europe (in alphabetical order)[8]

SiBol_93			*SiBol_05*		
Keyword	Freq. 93	Freq. 05	Keyword	Freq. 05	Freq. 93
brussels	6247	4619	barroso	562	6
ec's	1067	27	constitution	7577	4059
eec	1128	140	eu	20197	746
emu	291	97	euro	9508	4570
erm	4685	554	euronext	1154	0
erm's	99	1	euros	4188	4
europe	26715	31975	eurosceptic	463	48
european	40625	44532	eurosport	654	47
eurotunnel	932	684	eurostar	902	99
maastricht	9169	301	eurozone	1678	0
maastricht's	104	2	eurozone's	155	0
monetary	4890	2470	eu's	1770	55
ratification	1189	395			
ratify	743	239			
ratifying	231	54			
referendum	4639	3776			
treaty	7837	2888			
union	26415	21784			

The threshold for 'keyness' was set at a low p value of $p \leq 10^{-15}$in order to limit the results to the most statistically significant. The full keyword lists were then manually (and therefore subjectively) scanned for items which were considered to relate to the European Union and these items are shown in table 1 above.

As would be expected, several items are clearly "seasonal"[9] and affected by local events, such as *Maastricht* in *SiBol_93* compared to *constitution* in *SiBol_05*, or alternatively the frequency of *Eurotunnel* in *SiBol_93* when the channel tunnel was still under construction, compared to the frequency of *Eurostar* in *SiBol_05* when the channel tunnel was open and in use.

4.3 The decreasing visibility or newsworthiness of Europe and the European Union

The most striking feature of the keyword list is the significant decrease in the amount of attention that Europe receives in the newspapers, with keywords for *SiBol_93* encompassing *Europe*, *European*, *Brussels* and *Union.* Furthermore, the keywords list also shows that the newspapers in *SiBol_93* appear more interested in, or more willing to give space to, the process of the treaties, as keywords included *ratify*, *ratifying*, *ratification*, *treaty*, *referendum.* This is also of interest given that, in an earlier study contrasting explicitly anti- and pro-EU discourses (Taylor 2005) it was noted that a focus on the processes of the union was a feature of pro-EU texts.

This decreasing newsworthiness of Europe and its processes was also borne out by a further quantitative calculation of the percentage of articles referring to *Europe/European/EC/EEC/EU* in the two time periods, as shown in table 2.

Table 2: Percentage of articles in the sub-corpora containing *Europe/European/EEC/EC/EU*

	SiBol_93	*SiBo_l05*
Guardian	48.97	35.12
Daily Telegraph	34.95	33.66
Sunday Telegraph	45.32	40.69
Times	38.59	30.27
Sunday Times	43.21	29.23
Total	40.95	32.48

In each case, a clear decrease is evident, although the extent of this varies according to the newspaper: the *Sunday Times* shows the largest change in mentions of Europe, while the *Daily Telegraph* shows only a slight decrease. Overall, there is a decrease of 20 percent from 1993 to 2005 in the number of articles mentioning Europe.

An analysis of mentions of the two treaties and the ways in which the papers referred to them also confirms the trend seen above of the decreasing newsworthiness or visibility of Europe.

Table 3: Frequency of references to the two treaties, and ways of referring to them, in the year of ratification

	SiBol_93	*SiBol_05*
Treaty on European Union	86	6
Maastricht Treaty	2381	132
Treaty of Maastricht	49	6
European Constitution	3	736
Treaty Establishing a Constitution for Europe	0	11
Constitution for Europe	2	20
Treaty Establishing a European Constitution	0	2
Treaty Establishing a Constitution for the European Union	0	16

As noted in the introduction, in both cases the official names of the treaties were not the dominant way of referring to them and the frequencies of the variations can be seen in table 3 above. It is perhaps indicative of the attitude taken in 2005, that in *SiBol_05* the dominant term employed is *European Constitution*, therefore not even making reference to the fact that it was actually a treaty which was being proposed.

However, the most salient finding in table 3 is the frequency of *Maastricht Treaty* in *SiBol_93* (2381 occurrences, 24 per million words or pmw) compared to *European Constitution* in *SiBol_05* (736 occurrences, five pmw).[10] Evidently, in *SiBol_05* there was far less reporting of the treaty, which is also in line with our previous findings that Europe in general appears to be less newsworthy in *SiBol_05*. In view of this, it is also revealing to note how very few occurrences there are in *SiBol_05* of *Treaty Establishing a Constitution for Europe*, the official name of the treaty. It may be that the official name was somewhat unwieldy for the tightly written newspaper articles but just eleven occurrences in the year of ratification is remarkably few. Of the phrase *Treaty establishing a constitution for the European Union* all cases were in the form of the question *Should the UK approve the treaty establishing a constitution for the European Union?* which was proposed as the UK referendum question.

4.4 Establishing the EU

Returning to our keywords (see table 1), we can also begin to trace the linguistic establishment of Europe. For instance, in *SiBol_93*, the year in which it was established, *European Union* is the preferred form (1217 occurrences), which contrasts with *SiBol_05*, when *European Union* has been superseded by *EU* (3437 compared to 20,197 occurrences). It is also noticeable that the process of *European union* is far more salient in 1993 compared to the finished product of

European Union in 2005 (357 instances written with lower case in *SiBol_93* compared to just 21 in the much larger *SiBol_05*).

The term *Eurosceptic*, which appeared with a keyness value of 247 in *SiBol_05* compared to the reference corpus of *SiBol_93* (see table 1), is an example of language change rather than an increase in the reporting of anti-EU attitudes, as table 4 shows.

Table 4: Comparison of word forms in *SiBol_93* and *SiBol_05*

	SiBol_93	*SiBol_05*
euro-sceptic / Euro-sceptic	362	65
euro-sceptics / Euro-sceptics	728	33
eurosceptic / Eurosceptic	46	496
eurosceptics / Eurosceptics	112	232

As can be seen from table 4, although both the forms *Eurosceptic* and *Euro-sceptic* are found in *SiBol_93*, the dominant form is hyphenated (*Euro-sceptic*), while the opposite is true for *SiBol_05 (Eurosceptic)*. Extending the analysis showed that this pattern was repeated on a larger scale through the use of hyphens in all *Euro-* words. In *SiBol_93* there were 3106 instances of *Euro-*/*euro-** compared to just 809 in *SiBol_05*. The dropping of the hyphen indicates an acceptance and familiarity with the term, it has become established.

Furthermore, in *SiBol_93* not only are there more words starting with the head *euro** (80 pmw vs. 60 pmw), but there is also more creativity in the words formed from *euro**: there were 1125 different *euro** word types in *SiBol_93*, compared to 760 in *SiBol_05*. These examples illustrate the ways in which, as Europe and the European Union become 'established', the lexical resources for talking about it consolidate.

However, it is perhaps more revealing to look at the differences between the papers: as table 5 shows, the *Telegraph* is clearly the most linguistically resistant to the 'establishment' of Europe.

Table 5: Occurrences of *euro-** in the *Guardian*, *Telegraph* and *Times*[11]

	SiBol_93		*SiBol_05*	
	freq.	pmw	freq.	pmw
Guardian	806	8.1	150	1.0
Telegraph	1305	13.2	446	3.1
Times	995	10.1	213	1.5
Total	3106	31.4	809	5.5

Although all three newspapers show a decrease in frequency in 2005 compared to 1993, the relative frequencies are consistent, and in 2005 the occurrence of *euro** words in the *Telegraph* are three times more frequent than in the *Guardian*, and

twice as common as in the *Times*. As the *Telegraph Style Book*, its official guide to house style, itself states "compound words increasingly lose their hyphens as they are accepted as normal usage", therefore, we can conclude that the preference for *euro* followed by a hyphen in part indicates a distancing rhetoric in which the concepts of Europe and European Union are not accepted or normalised.[12]

4.5 Euro-* evaluation

The frequency of *Euro-**, as discussed in the section above, is much lower in *SiBol_05* compared to *SiBol_93*, and most frequent in the *Telegraph*. In both *SiBol_93* and *SiBol_05*, the *Telegraph* showed the greatest creativity in *Euro-** word forms. Intuition, reading the newspapers in our corpus, and previous research (Taylor 2005) which found that creative *euro** compounds were more frequent in an anti-EU corpus compared to a pro-EU corpus, suggested that the continued usage of *Euro** in *SiBol_05* was also likely to be used to negatively evaluate. This is illustrated in the *Telegraph* editorial extract below, which appeared under the heading "Mere democracy won't stop the EU machine":[13]

(13) In defiance of a united media, a monolithically pro-Brussels political class and blizzards of propaganda, they have said a resounding Non to the *Euro-elites* who have governed them for half a century. (*Telegraph*, 30/05/2005)

Other variations from the *Telegraph* in 2005 include the following, (in addition to *eurocrats* in examples (1) and (8) above):

(14) Perhaps it is a mercy that Mr Cook who, for all his *Euro-enthusiasm*, was a great parliamentarian, will not be around to witness the slow eclipse of the institution he admired. (*Telegraph*, 08/08/2005)

(15) Now, *Eurocrats* will first have to determine whether their ideas will bankrupt swathes of commerce and industry before turning them into directives. Welcome to planet Earth at last. Mr Verheugen reckons past efforts to halt the regulation juggernaut failed because of a lack of "political" backing, an understated way of saying they were torn to shreds by dirigiste governments, Left-wing *Euro-MP*s, and the army of lobbyists in Brussels. (*Telegraph*, 17/03/2005)

It could, however, be argued that in the *Telegraph* any terms connected with the European Union are likely to be used negatively because the newspaper's editorial policy is anti-EU. Therefore, to test this primarily subjective idea, rather than looking for corroboration, we also looked at the occurrences of *Euro-** in the *Guardian*, which, as the analysis of the editorials showed, is explicitly in favour of the European Union. The close analysis of the extended concordance showed that creative *Euro-** words were clearly evaluative in over half of the 87 instances, in particular with reference to music and cinema, and the evaluation

was overwhelmingly negative. Interestingly, several of the negative examples co-occurred with references to Conservative entities (in italics), as illustrated in examples (16), (17) and (18), suggesting an echoic mention of the anti-EU discourse.

(16) Those blue-sky thinkers just have more vision, shadow chancellor, so the baton has been passed to Graham Brady, *the shadow minister for Europe*, who dreams of the UK ditching those *ghastly Euro-laws* and adopting legislation from much further afield. Graham was debating at the London School of Economics on Monday when he was asked to put some bones on his party's plan to withdraw from the common fisheries policy. Hardly a thriller, until Graham said that once we're out, we'd copy the rules from places like the Falkland Islands. That's a cracking role model (and punchline) we tell him. (*Guardian*, 23/03/2005)

(17) The *Euro-headbangers* at the *Daily Telegraph* had thought to mark the occasion by printing minutes of an Anglo-French summit in 1971 at which Sir Edward, as prime minister, promised to abandon the US and take sterling into a single currency if Paris would say Oui. (*Guardian*, 26/07/2005)

(18) But he had a project – just as Tony Blair seemed to have a project – which was burying away, for the foreseeable future, a *Conservative party* trapped in *Euro-delirium* and patriotic fantasising. (*Guardian*, 18/04/2005)

This introspection-driven idea that *Euro** is likely to be used evaluatively also seems to be partially confirmed by the ways of referring to MPs in the European Parliament (see example (14)). As table 6 shows, all three broadsheets used the officially recognised form *MEP*, but only the *Telegraph*, which was openly anti-EU, showed a marked use of *euro-mp**. This seems significant because, in this case there is a selection of terms from which to choose.

Table 6: Frequency of references to *MEP/s*, *Euro-MP/s* and *EU-MP/s* in the three papers

	gua05	*tim05*	*tel05*
euro-mp/s / Euro-mp/s	3	5	62
EU-mp/s	0	0	0
MEP/s	416	540	450

By rejecting the official accepted version, *MEP*, it may be argued that the discourse of the *Telegraph* distances and delegitimises the MPs by denying them their preferred, or proper, title. This generalising function of the *euro* prefix may be seen to work, rather like the metonymic use of *Brussels*, to persuasively background the plurality of institutions and foreground a dominant foreign entity.[14]

4.6 Euro* - headlines

Frequency is not always a good indicator of relevance, in terms of mainstream discourse(s), some news types are more important than others and so are some places within the newspaper and some positions within news items. It is a long-standing rule of journalism to have "important elements at the beginning, less important at the end" (Mencher 1994: 105) and all news begins with a headline. Headlines have a very important role in news making and understanding. They summarise news, they select what is important thus establishing a hierarchy and indicating the path the story will take. In the Anglo-Saxon journalistic tradition, the headline and first paragraph constitute the 'lead' of a news item (Bell 1991), i.e. the textual unit containing the key elements and information that will be developed in the article. Since headlines are a clear signal of the focus of a story and are meant to guide interpretation, they seemed a perfect area in order to study patterns of representation and, more specifically, the attitudes of the newspapers towards Europe.

Table 7: *SiBol_93* keywords for Europe-related headlines

SiBol_93 Keyword	Freq.	%	RC. Freq.	RC. %	Keyness
city	283	1.00	28	0.06	400.19
business	266	0.94	71	0.15	247.15
steel	59	0.21	1		107.75
focus	92	0.33	16	0.03	106.71
maastricht	34	0.12	0		67.40
gatt	34	0.12	0		67.40
erm	31	0.11	0		61.46
clinton	29	0.10	0		57.49
major	46	0.16	7	0.01	56.32
delors	28	0.10	0		55.51
watch	37	0.13	3		54.83
copenhagen	23	0.08	0		45.59
kiosk	20	0.07	0		39.64
bank	49	0.17	19	0.04	34.20
monetary	17	0.06	0		33.70
japan	19	0.07	1		30.65
car	37	0.13	12	0.03	29.93
recession	14	0.05	0		27.75
hurd	13	0.05	0		25.77
ministers	34	0.12	12	0.03	25.73
single	35	0.12	13	0.03	25.37

Table 8: *SiBol_05* keywords for Europe-related headlines

SiBol_05 Keyword	Freq.	%	RC. Freq.	RC. %	Keyness
blair	191	0.40	1		167.01
constitution	170	0.36	2		140.16
rebate	118	0.25	0		109.62
comment	164	0.34	20	0.07	65.46
analysis	112	0.23	8	0.03	61.09
budget	107	0.22	7	0.02	60.62
chirac	61	0.13	1		48.38
turkey	64	0.13	3	0.01	40.87
china	53	0.11	2		35.99
blair's	30	0.06	0		27.85
referendum	58	0.12	6	0.02	25.91
brown	41	0.09	2		25.85
mandelson	27	0.06	0		25.06
world	127	0.27	30	0.11	24.19
bush	26	0.05	0		24.13
sugar	26	0.05	0		24.13

In order to access articles where Europe is the main subject we narrowed down the query to headlines containing the following search terms: *europ**, *EU*, *EC* or *EEC*. Headlines and lead paragraphs have been annotated in the corpus and are easily retrievable using Xaira's XML query tool. We extracted Europe-related headlines for the two corpora and generated keywords, comparing them against each other (tables 7 and 8 above).

In the keywords list referring to 1993 (table 7) we see a prevalence of words relating to the business/economics sphere. Some are explicitly business terms (e.g. *monetary*, *GATT*, *recession*, etc.), some signal business/financial sections of the newspaper (i.e. *City*, *Focus*, *Watch*, *Kiosk*), others turn out to refer to the economy when the enlarged context of concordance lines is looked at (e.g. *steel*, *Japan*, *car*).

In the 2005 list (table 8), the 'business' semantic field is still there (e.g. *budget*, *China*, *sugar*), but there is a much larger presence of words belonging to the political (national, European and international) sphere. We find names of national/international politicians, words relating to contingent political events (e.g. *constitution*, *treaty*) and parts of the newspapers indicating political analysis (i.e. *Comment & Analysis*, which refers to the columns' page).

Collocation analysis, using Xaira software, confirmed that in *SiBol_93* news about Europe mainly concerned sport and economics, while the political Europe seemed to be a minor topic. The European Union was established as a political entity in 1993, but neither *European Union* nor the acronym *EU* are commonly used in the press. The term *European Community* (or *EC/EEC*)

making reference to the economic entity (the common market) founded in 1957, is preferred and more established. In the 1993 headlines, there are only 44 occurrences of *EU/European Union*, 32 of which are in the *Guardian*, while *European Community/EC/ECC* is mentioned 1054 times. As noted in the analysis of the editorials, the choice of referring to the institutional Europe as *EC* is not casual, but reflects a specific position of the newspaper.

Albeit in limited proportion, the *Guardian* breaks new ground in mentioning the EU and in the coverage of the newly born political institution. In order to give an idea of what was in the news and also to give a sense of proportions, all the EU-related headlines that appeared in the *Guardian* in 1993 are listed below (19 to 50).

(19) EU agrees targets for recycling (*Guardian*, 27/12/1993)
(20) EU and US strike farm trade deal (*Guardian*, 07/12/1993)
(21) British Steel in plea to EU as it claws way back to profit (*Guardian*, 16/11/1993)
(22) Norwegians risk missing EU boat (*Guardian*, 20/12/1993)
(23) A shift in common ground as European Union comes of age (*Guardian*, 01/11/1993)
(24) European Union criticised over fraud scandals (*Guardian*, 17/11/1993)
(25) EU draws back from sanctions pledge to Serbia (*Guardian*, 18/11/1993)
(26) EU expansion in trouble over terms of membership (*Guardian*, 24/11/1993)
(27) Sweden in a huff as EU gets sniffy over favourite snuff (*Guardian*, 24/11/1993)
(28) Britain moves to use up EU grants (*Guardian*, 20/11/1993)
(29) News in brief: EU inflation dips (*Guardian*, 21/12/1993)
(30) Commentary: European Union is in the eye of the beholders (*Guardian*, 02/11/1993)
(31) EC it may be, but EU it will be, if the meaning is clear (*Guardian*, 09/11/1993)
(32) Chancellor rejects EU jobs plan (*Guardian*, 23/11/1993)
(33) EU leaders to keep Greeks on tight rein (*Guardian*, 30/12/1993)
(34) European union of churches 'will fight on social issues' (*Guardian*, 18/11/1993)
(35) News in brief: Stasi spy 'in EU offices' (*Guardian*, 03/12/1993)
(36) Sinn Fein urges European Union officials to step up pressure on Major (*Guardian*, 07/12/1993)
(37) Hunt blocks EU paternity move (*Guardian*, 24/11/1993)
(38) EU reaction: Door is opened on lucrative markets (*Guardian*, 15/12/1993)
(39) Ministers spurn pounds 33bn EU recovery plan (*Guardian*, 06/12/1993)
(40) EU says states can pass seat-belt laws (*Guardian*, 12/11/1993)
(41) EU states forge Macedonia ties (*Guardian*, 17/12/1993)
(42) EU states plan links with Macedonia (*Guardian*, 30/11/1993)
(43) Ministers back EU steel rescue plan (*Guardian*, 18/12/1993)

(44) EU summit moves focus off jobs (*Guardian*, 07/12/1993)
(45) EU support launches Delors jobs scheme (*Guardian*, 13/12/1993)
(46) EU talks raise hopes of Greek and Turkish help on Cyprus (*Guardian*, 10/11/1993)
(47) EU to recognise Macedonia (*Guardian*, 15/12/1993)
(48) Birt tells EU to avoid TV Euroculture (*Guardian*, 01/12/1993)
(49) EU urged to embrace east Europe (*Guardian*, 21/12/1993)
(50) EU warns Major over 'opt-out' (*Guardian*, 15/12/1993)

In the *Guardian* headlines, the EU is represented as something that exists and acts (e.g. it "says", "warns", "urges", "moves", "launches", "reacts", "forges" ...) and whose actions produce effects. The evaluation is varied, and in most cases not explicit, this because nearly all the headlines describe hard news items. What is also interesting is the fact that the *Guardian* talks about the EU in news reports and not just in comment articles. This does not happen in the other newspapers, certainly not in the *Telegraph* which only has two headlines dealing with the EU (an editorial in the *Telegraph* and an opinion-editorial in the *Sunday Telegraph*), which are shown in (51) and (52).

(51) Sunday Comment: Welcome to the EU – you might not be there long (*Telegraph*, 31/10/1993)
(52) Leading Article: EU, eh? (*Telegraph*, 03/10/1993)

What would make the headlines about the EU in the *Telegraph* is its absence, the newspaper does not acknowledge the recently established Union. Furthermore, since the birth of the EU is a 1993 story, the fact that the *EU* does not 'grab' a single news-headline in the *Telegraph* could reasonably be interpreted as careful avoidance.

The *Times* gives a little more space to the European Union, in order to criticise it (examples (53)-(58)). In all the headlines mentioning the EU, negative evaluation is quite evident. The EU is presented as a "threat" to the economy (55) and a place of sleazy controversy (it is "born covered in confusion" – (56)), it sneaks in "with little fanfare" (57), arriving unwanted and unplanned (it "slips into being" – (57)) and producing immature babble ("sends confused signal" – (58)).

(53) UK's economic ills make France *fear* for European union (*Times*, 06/02/1993)
(54) Bad week for European Union (*Times*, 18/04/1993)
(55) European union *is a threat* to world trade (Business Letter *Times*, 18/04/1993)
(56) European Union *is born* covered *in confusion* (*Times*, 01/11/193)
(57) European Union *slips into being* with *little fanfare* (*Times*, 02/11/1993)
(58) European Union *sends confused signal* to the warring factions (Commentary; Balkans *Times*, 24/11/1993)

Headlines in 2005 project a different picture of Europe. As we have seen from the keywords, a political pattern starts to emerge and take over the economic one. Similar findings are confirmed by retrieving collocations using Xaira. Sport continues to be at the top of the hierarchy of Europe's newsworthiness but the political EU gains position. In *SiBol_05*'s headlines there are 1327 occurrences of *EU/European Union*, quite evenly distributed across newspapers. The *European Community/EC/EEC* is still there, but the proportion has inverted and we only find a total of 47 occurrences. When normalising frequencies to the corpus size, the presence of *EU/European Union* in the headlines increases from 0.46 per million words in 1993 to 9.28 per million words in 2005.

A large portion of the headlines concerns the European constitution. We looked in detail at the subset of headlines focussing on the constitution, in order to see what was the individual paper's stance towards the topic and to get an idea of change/consistency of attitudes towards the EU. The diverging positions of the three newspapers are exemplified by the following examples ((59) to (86)), the examples included are a sample of the 142 headlines explicitly dealing with the constitutional referendums and have been chosen for their representativeness of the general attitudes of the *Guardian*, the *Telegraph* and the *Times*, respectively.

The defeat scenario of the referendum on the European constitution is represented in the *Guardian* in terms of danger and crisis. Europe is under threat, "on a knife edge", "on ice", hopeless efforts are made in order to save it, but the perspective is one of uncontrollable disaster (see examples: (66)-(67)). The *Guardian* also focuses on the aftermath, extensively using metaphors of difficulty and distance (see examples: (68)-(71)).

(59) Chirac *seeks to salvage* EU constitution (*Guardian*, 13/04/2005)
(60) Schroder comes to *aid* of Chirac's campaign for EU constitution (*Guardian*, 27/04/2005)
(61) Dutch vote *could kill off* EU blueprint (*Guardian*, 31/05/2005)
(62) French voters *threaten* to shun EU treaty (*Guardian*, 19/03/2005)
(63) Comment & Analysis: EU constitution: Europe *on a knife edge* (*Guardian*, 23/05/2005)
(64) Blair wins budget allies as EU summit puts constitution *on ice* (*Guardian*, 17/06/2005)
(65) France *braces* for no vote *fallout*: Prospect of heavy defeat in vote on constitution signals gravest *crisis* in European project's history (*Guardian*, 28/05/2005)
(66) EU referendum prompts French identity *crisis* (*Guardian*, 24/05/2005)
(67) UK diplomat criticises '*cavalry charge to disaster*' over EU constitution (*Guardian*, 05/09/2005)
(68) Comment & Analysis: Think about life after rejection: The EU referendum isn't *a glib opportunity for scoring points* (*Guardian*, 09/05/2005)
(69) Comment & Analysis: French EU referendum: *Try harder*, Jacques (*Guardian*, 16/04/2005)

(70) Comment & Analysis: The French referendum shows *how far old Europe has travelled* (*Guardian*, 30/05/2005)
(71) French referendum: EU leaders express regrets and prepare to *play long game* (*Guardian*, 30/05/2005)

The *Telegraph* confirms its 'euroscepticism' by stressing the blurriness of the EU constitution, which the *Telegraph* presents as being vague (72), dull (73), equivocal (75), not exactly democratic (76) and (77) and as causing indecision (74). Negative evaluation is quite overt also in the headlines focussing on the 'Yes' campaign (a "stuttering campaign" and a "plea"), depicted as a 'lost cause' (see examples: (78)-(79)).

(72) *Less fudge, more facts* about EU constitution (*Telegraph*, 31/01/2005)
(73) Passion? Vision? *Perhaps* the EU referendum will be *more lively* (*Telegraph*, 15/04/2005)
(74) THE EUROPEAN QUESTION EU constitution? *Yes, but no, but yes, but no...* (*Telegraph*, 22/05/2005)
(75) Unless the Tories *find* themselves, Blair could win the EU referendum (*Telegraph*, 10/01/2005)
(76) Spain *gears up* for EU referendum (*Telegraph*, 12/02/2005)
(77) EU Constitution *must have support of majority* (*Telegraph*, 21/05/2005)
(78) Is Chirac *too late to save* EU constitution? Nervous president joins the *stuttering campaign* for a Yes vote tonight but opponents scent *a chance* for the French public to kill draft treaty stone dead (*Telegraph*, 14/04/2005)
(79) Chirac *plea* for Yes vote on EU treaty (*Telegraph*, 29/03/2005)

The *Times* presents the EU constitution as a marketing operation, something that needs to be sexed up and sold or sneaked in before the people ("they") realise (see examples: (80)-(82)).

(80) They don't like the EU constitution? *Quick, send in the force* (*Times*, 21/01/2005)
(81) Chirac begins *hard sell* on EU constitution (*Times*, 01/03/2005)
(82) EU referendum question *made easy on the eye* (*Times*, 27/01/2005)

The *Times* also focuses extensively on the failure of the referendum as the defeat of the EU and its promoters (see examples: (83)-(86)).

(83) Constitution leaves EU *weak* and *broken* (*Times*, 17/06/2005)
(84) Dreams of a bigger EU *dashed* by voters' fears for lost jobs (*Times*, 01/06/2005)
(85) Chirac Cabinet *tears itself apart* over EU constitution (*Times*, 19/04/2005)
(86) French *in disarray* as they admit EU treaty vote is lost (*Times*, 26/05/2005)

In 2005, the political Europe is an established reality, it gets similar quantitative coverage in all the newspapers under investigation, while in *SiBol_93* the frequencies were low and the distribution disproportioned. This indicates a change in the status of the EU, but what seems to have remained constant is the attitude of the three main broadsheets towards the EU. Confirming our other findings, the papers tending towards the conservative side of the political spectrum (*Telegraph* and *Times*) are characterised by a marked antagonism towards the EU while the *Guardian* tends to be more informative and have a generally friendly attitude.

5. Collocation analysis

The collocation analysis of relevant items such as *Europe*, *European*, *Brussels* etc. revealed that the main difference in the representation of Europe was quantitative rather than qualitative. As noted above, there was a sharp decrease in the amount of coverage allowed to European issues in *SiBol_05* compared to *SiBol_93*, but a comparison of the collocates for the two years showed a high degree of consistency.

Following the thread of the previous headline analysis, we also compared the European Community of *SiBol_93* with the European Union of *SiBol_05*. The comparison could not be based on the same definition of institutional Europe in the two periods, because, as has been pointed out, *EU*/*European Union* in 1993 had a marginal presence and a different status.

We examined the corpus in order to understand whether, and how, the meaning of institutional Europe has changed with the change of its labelling. The starting point was a collocation analysis of the two sets of lexical items, namely *EC/EEC/European Community* in *SiBol_93* and *EU/European Union* in *SiBol_05*.

Among the top collocates (L5 R5) of *EC/EEC/European Community* in *SiBol_93* (figure 2) we find *countries* with 994 occurrences (and a z-score=180.2), 79 percent of the cases refer to European countries.[15] Further down in the list we find *states* (431 occurrences and a z-score=68.9), a closer analysis of the concordance lines reveals that a third of the occurrences of *states* refer to the *United States* and two thirds are references to member states.

The situation changes in *SiBol_05* (figure 3). *Countries* remains a top collocate for *EU/European Union* with 945 occurrences (and z-score=150.0), but *states* gains several positions in the list of collocates with 809 occurrences (z-score=151.6), 88 percent of which refer to member states. The reference to member states has passed from 3.0 per million words in *SiBol_93* to 4.8 per million words in *SiBol_05*.

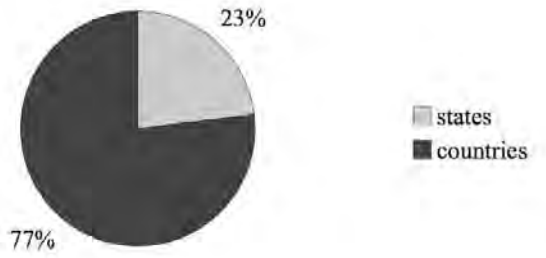

Figure 2: The proportion *state-country* in the affiliation to the EC (*SiBol_93*)

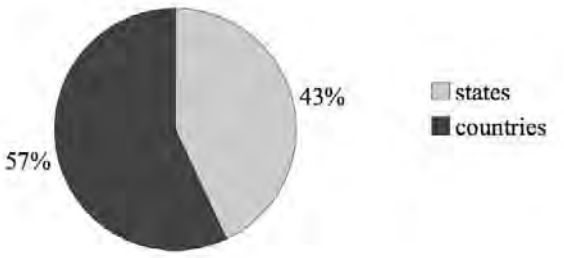

Figure 3: The proportion *state-country* in the affiliation to the EU (*SiBol_05*)

This shift in the 'preferred' collocates may suggest that the idea of a federal Europe of states has penetrated the press. There is probably not enough evidence to conclude that the newspapers' discourse is accompanying the morphing of the federal Europe, but the federal image is there. We cannot say that there has been a change in the sense of belonging, but there has been a movement in the formal definition of membership. The interpretation of this phenomenon is not unproblematic and far too complex to be limited to a linguistic point of view, as it involves philosophical, historical, political and social elements. Our analysis does not aim at giving a comprehensive explanation of the process of European integration (or Europeanisation), but we claim that the textual analysis of newspaper discourse "might help us understand what Europe means" (Millar and Wilson 2007: 1).

6. Conclusions

In this paper we have attempted to briefly outline some of the ways in which Europe has become established in the UK press, and how the representation of Europe and the European Union changed from the inception of the Union in 1993 to 2005, the year in which the treaty on a European Constitution was proposed. We have also identified a set of areas that may prove fruitful for further research. The most salient finding is not so much the negativity regarding the construal of Europe, which was overtly signalled, but the decreasing visibility of Europe, and we consider this to have both methodological and social implications. First, the methodological implications lie in showing how corpus linguistics, when used comparatively, can actually identify the absent as well as the frequent. Second, in terms of social implication, de Vreese et al. (2006: 478) conclude that "negative news and, in general, a lack of news regarding the EU and EP [European Parliament] are thought to contribute to a lack of legitimacy and to detract from the formation of a European identity". Our research indicates that collocates of Europe-related terms remained relatively stable over the twelve year period, and the main issue that the European Union faces is one of newsworthiness and visibility.

Notes

1 The European Commission's Press Office in London currently maintains the following webpage dedicated to monitoring "the British press's highly distorted coverage of the European Union" and reporting on the "euro-myths" which the papers carry: http://ec.europa.eu/unitedkingdom/press/euromyths/index_en.htm.

2 Integrated and United? A Quest for Citizenship in an 'Ever Closer Europe' (*IntUne*) is a four year project financed by the European Union within the scope of the 6th Framework Programme www.intune.it.

3 eXtensible Mark-up Language (http://www.w3.org/XML/).

4 Text Encoding Initiative (http://www.tei-c.org/index.xml).

5 Xaira (XML Aware Indexing and Retrieval Architecture) developed at Oxford University is the XML version of Sara, the software originally developed for interrogating the *British National Corpus*. For further information: http://www.oucs.ox.ac.uk/rts/xaira/.

6 See WordSmith 5.0 manual at: http://www.lexically.net/downloads/version5/HTML/index.html.

7 Here and elsewhere, italics added for emphasis.

8 Word frequencies used to calculate keywords were not case sensitive.

9 Gabrielatos and Baker (2006) use the term 'seasonal collocates' in contrast with 'consistent collocates' which are items that remain stable over a time period and are not influenced by local events.

10 To allow for comparison of relative frequencies, the occurrences have been normalised to per million words (pmw).

11 Including *Sunday Times* and *Sunday Telegraph.*

12 http://www.telegraph.co.uk/news/1435295/*Telegraph*-Style-Book-Introduction.html.

13 There were four other *Euro-** words in this editorial adding to the cumulative impression of 'euro-dismissal'.

14 As would be expected, *Brussels* was most frequent in the *Telegraph* in 2005, and the 15 collocates with the highest z-score included the terms; *eurocrats*, *bureaucrats*, *unelected*, *emanating* (only *bureaucrats* appeared in the same list for the *Guardian*).

15 The z-score is a measure of collocational significance. Xaira allows the selection of either z-score or MI (mutual information). The z-score indicates standard deviation from the mean frequency and describes the probability that the node and the collocate are related: the higher the z-score, the higher the degree of collocability of the node word with the item.

References

Anderson, J. (2004), 'A flag of convenience? Discourse and motivations of the London-based Eurosceptic press', *European Studies*, 20: 151-170.

Baker, P. (2006), *Using Corpora in Discourse Analysis.* London: Continuum.

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyżanowski, T. McEnery and R. Wodak (2008), 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', *Discourse and Society*, 19(3): 273-305.

Bell, A. (1991), *The Language of News Media*. Oxford: Blackwell.

Firmstone, J. (2003), ''Britain in the Euro?': British newspaper editorial coverage of the introduction of the Euro', *European Political Communication Working Paper Series*, 5(3).

Gabrielatos, C. and P. Baker (2006), 'Representation of refugees and asylum seekers in UK newspapers: towards a corpus based analysis'. Paper presented at the Joint Annual Meeting of the British Association for

Applied Linguistics and the Irish Association for Applied Linguistics (BAAL/IRAAL): *From Applied Linguistics to Linguistics Applied: Issues, Practices, Trends*, 7-9 September 2006, University College, Cork, Ireland.

Hardt-Mautner, G. (1995), ''Only connect'. Critical discourse analysis and corpus linguistics'. *UCREL Technical Paper 6*. Lancaster: University of Lancaster.

Mencher, M. (1994), *News Reporting and Writing* (6th edition). Madison and Dubuque: Brown and Benchmark.

Millar, S. and J. Wilson (eds.) (2007), *The Discourse of Europe: Talk and Text in Everyday Life.* Amsterdam: John Benjamins.

Morley, J. (2004), 'The sting in the tail: persuasion in English editorial discourse', in: A. Partington, J. Morley and L. Haarman (eds.) *Corpora and Discourse*. Berlin: Peter Lang. 233-247.

Partington, A. (2004), 'Corpora and discourse: a most congruous beast', in: A. Partington, J. Morley and L. Haarman (eds.) *Corpora and Discourse.* Bern: Peter Lang. 11-20.

Partington, A. (2006), 'Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work', in: A. Stefanowitsch and S.Th. Gries (eds.) *Corpus-based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter. 267-304.

Partington, A. (forthcoming), 'Evaluating evaluation and some concluding thoughts on CADS', in: J. Morley and P. Bayley (eds.) *Corpus-Assisted Discourse Studies on the War in Iraq*. London: Routledge.

Partington, A. and A. Duguid (2008), 'Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS)', in: M. Bertuccelli Papi, A. Bertacca and S. Bruti (eds.) *Threads in the Complex Fabric of Language*. Pisa: Felici Editore. 271- 280.

Peter, J. and C.H. de Vreese (2004), 'In search of Europe: a cross-national comparative study of the European Union in national television news', *Press/Politics*, 9(4): 3-24.

Scott, M. (2008), *WordSmith Tools*, version 5, Liverpool: Lexical Analysis Software.

Stubbs, M. (2001), *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Taylor, C. (2005), 'Electronic corpora in language study', *Dafwerkstatt*, 6: 71-84.

Vreese, C.H. de, S.A. Banducci, H.A. Semetko, and H.G. Boomgaarden (2006), 'The news coverage of the 2004 European parliamentary election campaign in 25 countries', *European Union Politics*, 7(4): 477-504.

Part II

Lexis, grammar and semantics

A nightmare of a trip, *a gem of a hotel*: The study of an evaluative and descriptive frame

Stephen Coffey

University of Pisa

Abstract

This paper describes a lexico-grammatical frame which may be exemplified by the phrase a nightmare of a trip*. The internal semantics are a defining feature of the frame: the second noun is likened in some way to the first noun (the 'trip' was a 'nightmare'). By examining many sets of concordance lines generated from the* British National Corpus*, a database was created containing 380 examples of the frame. The present article describes these phrases from a number of points of view: their internal lexico-semantic structure, their communicative functions, their place within syntactic structure, and their relation to text type.*

1. Introduction

In this article I discuss phrases such as *a nightmare of a trip* and *a gem of a hotel*; description is based on examples found in the *British National Corpus* (hereafter BNC). The basic phrasal type is described in the *Oxford English Dictionary* (hereafter OED) at the entry for *of* (use no. 23, 'between two nouns which are in virtual apposition', subsection b. 'in the form of, in the guise of'). The OED states that, 'the leading noun is the latter, to which the preceding noun with *of* stands as a qualification, equivalent to an adjective; thus "that fool of a man" = that foolish man, that man who deserves to be called "fool"; "that beast of a place" = that beastly place'. Examples in the OED go back many centuries, e.g. 'That scamp of a husband of hers' (1850), 'The rogue of a clerk' (1791), 'That monster of a man, Lord Howard of Escrick' (1683).

Quirk et al. (1985: 1284-1285) briefly describe the phenomenon within a section entitled "apposition with *of*-phrases". They point out, among other things, that phrases such as 'an angel of a girl' and 'this jewel of an island' may be paraphrased as 'The girl is an angel / an angelic girl' and 'this island is a jewel / this jewel-like island'. They also note that the second noun must be preceded by *a/an*, and that the first noun must be singular ('?**those fools* of policemen' – see section 4.4 below). Some discussion from the point of view of 'intensification' is found in Bolinger (1972: 83-84).

Modifiers may precede either or both of the nouns, and the basic syntactic form of the frame may thus be described as '[*a*] (modifier) NOUN_1 *of a* (modifier) NOUN_2'. For the sake of simplicity, I will sometimes refer to the frame as the '*a*

NOUN$_1$ *of a* NOUN$_2$' frame; when discussing the first part of the expressions (e.g. 'a gem of a'), I will sometimes refer to the 'NOUN$_1$ phrase'.

There is, to my knowledge, little corpus-based description of the frame. Notably, there is no mention of it in the Longman grammar (Biber et al. 1999), presumably because it is not a very frequent frame, and is also quite variable in form. It does not emerge from Sinclair's exploratory study into the nature of *of* (1991: 85-98), but there is no reason why it should since the study was based on a relatively small number of corpus tokens of *of*. Similarly, it is absent from Meyer's (1992) corpus-based study of apposition. This may just be due to low frequency, especially since the partial corpora involved in this study amounted to only about 360,000 words of running text. In any case, the frame is not representative of the central notion of 'apposition', which involves a semantic relationship between different phrasal units and not within the same unit.

Two corpus-based works which do briefly mention the frame are Francis et al. (1998: 198), which contains a list of eight items comprising a sub-group (the 'giant' group) of the general pattern 'N *of* n', and Coffey (2008: 126), who mentions the phrasal type within a study of indefinite article usage.

2. Trawling the corpus

In order to find a reasonable number of realizations of the frame, various corpus searches were carried out. Some of these were for grammatical frames, specifically: *a _ of a/an*, *a _ _ of a/an*, *this _ of a/an*, *that _ of a/an*, and *one _ of a/an*. Where feasible, all tokens of a frame were retrieved in their KWIC contexts; where frequency was too high, a large random sample (several hundred tokens) was examined. At a later stage, the frame *the _ of a* was also investigated, but since it was much more frequent than the others (over 50,000 tokens), the sample was taken from a subcorpus of 'imaginative literature', where, it was discovered, the '*a* NOUN$_1$ *of a* NOUN$_2$' frame tends to occur more often. It turned out, however, that this search was not very useful.

An interesting finding with regard to the grammatical searches was that frequency of the syntactic frames was lower than had been expected (except in the case of *the _ of a*), and therefore it was easier, or at least less difficult, to find examples of the phenomenon being investigated. There are, for example, only 7,131 realizations in the BNC of the grammatical frame *a _ of a*. This is a very small number when compared with the slightly shorter frame *a _ of* and with the formally parallel *the _ of the*, for both of which there are over 250,000 corpus tokens.

In addition to examining these grammatical frames, some specific lexical items were looked for, mainly items which had been found during the grammatical searches, either in N_1 or in N_2 position. For example, searches were made for *brute of*, *a _ of a man*, and *humdinger*. The exact nature of the search depended on the corpus frequency of the individual word or string of words: that is, I wished to strike a happy medium between recall and precision, finding as

many pertinent examples as possible without the task becoming unacceptably labour-intensive.

Lexical searches were also carried out for potentially offensive words (e.g. *sod*), since it had been noted elsewhere (Coffey 2003) that words of this type were sometimes found in this frame. Finally, a small number of searches was made for phrases which I had noted in my own everyday contact with the language (e.g. 'a lash of a shot', heard in a television sports commentary).

3. Defining the data set

Before describing the phrases which were found in the corpus, a few points need to be made regarding items which, for one reason or another, were excluded from the study.

A first set of items are those which contain NOUN_1 phrases such as *a glimmer of a* (e.g. 'a glimmer of a smile'), *a ghost of a* and *a flicker of a.* Although these show similarities with the main body of phrases, they form a separate group that is quantitative in nature rather than descriptive or evaluative. They also have co-textual features which set them apart from others investigated, for example they are quite likely to be preceded by the definite article ('meeting her compassionate gaze with the glimmer of a smile'), and some of the NOUN_2s may be uncountable ('He looked briefly in my direction, but without a flicker of recognition').

A second group of items excluded from the study are a small set of phrases which are best considered as fully lexicalized units (or at least very restricted lexical frames). The phrases in question are: *a whale of a time*, *a broth of a boy*, *a slip of a [girl]* and *a [fine] figure of a man/woman*, (the latter with its atypical obligatory adjective). In the case of *a slip of a [girl]*, there is in fact a certain amount of variation in N_2 position (*a slip of a girl/child/lad/maid/thing*), but the various resulting phrases are best seen as constituting a single lexicalized frame, partly because of the similarity in meaning of the second nouns but especially because the word *slip* (in this meaning) was only found within the '*a* NOUN_1 *of a* NOUN_2' frame. *A slip of a [girl]* may be contrasted with a similar phrase, *a chit of a [girl]*, which was included because the word *chit* is also used on its own with the same meaning (e.g. 'How could the silly chit have so mistaken his intent?').

With regard to *a whale of a time*, it should be pointed out that other phrases were found which begin *a whale of a* ('a whale of a deal/hit/record/song/story/tale'), and these were included in the study.

In addition to the fixed phrases just discussed, there are also a few items in which one part of the frame is best considered as a fixed item. I have excluded phrases in which *a lot* constitutes the second part of the frame (as in 'She'd had *a dickens of a lot* to drink'). *A lot* is a fixed, two-word lexical unit and this fact in itself sets it aside from other N_2s. Also, it is preceded in all cases by an intensifer (*a devil of ...*, *a dickens of ...*, *a heck of ...*, *a hell of …* and *a fuck of ...*), and the

resulting phrases are probably best viewed as constituting a restricted lexical frame.

With regard to NOUN$_1$ phrases, I have excluded all items beginning with the phrase *a hell of a*. Frequency alone sets *a hell of a* apart from the rest: while the most frequent NOUN$_1$ phrase included in the data set occurs 25 times ('a gem of a'), there are a total of 510 corpus tokens of *a hell of a* in its various forms. Furthermore, the phrase has travelled very far from its origins as a descriptive item (N_2 is like 'hell'), and has become a 'general' modifier with a variety of functions. It can be used: (i) as an intensifier before noun phrases, including those with adjectives (e.g. 'it's a hell of a coincidence', 'that was a hell of a difficult one to find'); (ii) as an intensifier within the already cited *a hell of a lot*; (iii) as an evaluating pre-modifier of nouns (e.g. 'Rory can still put on one hell of a floorshow'), or within verbal expressions (e.g. 'He'll give you one hell of a time if you persist'). For the same reasons, it seemed appropriate to exclude the originally euphemistic alternative *a heck of a* (50 corpus tokens), which functions in a very similar way to *a hell of a*.

The total number of expressions included in the present study is 380.

4. The composition of phrases

In this section I describe the lexico-semantic composition of the phrases, with reference to the two nouns they contain, to any pre-modifiers, and to the first element in the double noun phrase, usually a determiner.

4.1 The first nouns: a *nightmare* of a trip

153 different items were found in N_1 position. These are of variable frequency, as can be seen in table 1.

Table 1: NOUN$_1$ type-token ratios

Tokens	1	2	3	4	5	6	7	8	9	11	12	19	20	25
Items	103	14	7	7	8	3	1	2	1	2	2	1	1	1

As with many other language phenomena, there is a tendency for the number of items to diminish as the number of tokens increases (Zipf's Law). In descending order of frequency, the items found more than six times are: *gem* (25), *giant* (20), *devil* (19), *bitch* (12), *brute* (12), *monster* (11), *peach* (11), *fool* (9), *bear* (8), *cracker* (8), *dream* (7).

The items with a frequency of between two and six are listed in table 2. The remaining items, those for which there was only one corpus token, are to be found within the complete list of nouns in figure 1.

Table 2: NOUN$_1$ items with frequency between two and six

Tokens	
6	bastard, jewel, whale
5	chit, dickens, humdinger, mountain, mouse, pig, slab, wreck
4	beast, beauty, nightmare, scrap, sod, swine, wisp
3	battleaxe, farce, hulk, rock, rocket, roller-coaster, scorcher
2	ass, beanpole, bugger, colossus, diamond, freak, gnome, joy, morgue, ponce, slut, sop, stormer, stump

---

The ANIMAL group: bear, beast, bird, bull-dog, butterfly, bullock, Dobermann, giraffe, hawk, hippopotamus, lioness, mongrel, mouse, pig, walrus, whale, whippet;
The OTHER CREATURES group: giant, gnome, monster, pixie, troll;
The PLANT group: ashpole, beanpole, reed, runner bean, stick;
The FOOD group: dumpling, peach;
The NATURAL PHENOMENA group: mountain, rock, whirlpool, whirlwind;
The PRECIOUS STONES group: diamond, gem, jewel, pearl;
The POSITIVE PEOPLE group: genius, grande dame, prodigy, wizard;
The NEGATIVE PEOPLE group: battleaxe, bitch, brute, burke, chit, eejit, fool, hag, humbug, pest, ponce, prat, rough diamond, scallywag, scamp, scoundrel, silly cow, sleeze-bag, slob, slut, swine, thief, villain, whore;
The PEOPLE group: godfather, grandmother, loner, maverick;
The VERY OFFENSIVE group: arseflap, ass, asshole, bastard, bugger, fuck, shit, shit-hole, sod;
The POSITIVE WORDS group: beauty, blockbuster, corker, cracker, dream, humdinger, joy, miracle, marvel, stunner:
The NEGATIVE WORDS group: abortion, dead duck, devil, dickens, disaster, farce, headache, joke, nightmare, ragbag, shambles, sop, wreck:
The DEATH group: crypt, grave, graveyard, morgue;
The MASS group: bundle, heap, hulk, hunk, jelly-heap, lump, shell, skeleton, slab;
The SPEED/STRENGTH group: belter, blitz, stormer, wizzer:
The SMALL SIZE group: knob, midget, scrag, scrap, sprout, stump, wisp;
The VEHICLES/MACHINES group: bulldozer, dynamo, piledriver, Porsche, rocket, roller coaster;
The MISCELLANEA group: bear hug, bolster, bone-shaker, bulwark, chop (movement), colossus, frown, freak, ham (the shape is important), head-thumper, iceblockbuster, papyrus, ponytail, razor, scorcher, sexquake, sink, tank-buster, trap, trickle, vocoder, whiplash.

Figure 1: Semantic groupings of NOUN$_1$ items

Many of the nouns in N_1 position may be grouped together on the basis of their core meaning (though actual text usage sometimes highlights different semantic features). Figure 1 shows the groups which suggest themselves, together with the individual lexical items (though some of the items could be placed in more than one group). Inevitably, the list ends with a somewhat mixed bag of 'miscellanea'. The list contains all nouns found in N_1 position.

4.2 The second nouns: a nightmare of a *trip*

A variety of nouns and of meanings was found in N_2 position. In two cases there is no second noun: that is, the phrases are incomplete. In a few other cases I have also excluded N_2 from the statistics below. This involved one phrase whose meaning is not clear to me, and eleven phrases in which it is the meaning of a verbal phrase which counts, and the single noun cannot be said to have autonomous meaning. An example of the latter is:

(1) I am having a devil of a time with a contract in Cirencester (novel)

The remaining total of NOUN$_2$s is 366. As was the case with N_1, the number of tokens is disproportionate to the number of items, as can be seen in table 3.

Table 3: NOUN$_2$ type-token ratios

Tokens	1	2	3	4	5	6	8	15	73
Items	155	18	13	4	2	1	2	1	1

The most frequent second nouns are *man* (73), *woman* (15), *girl* and *song* (8), *game* (6), *goal* and *thing* (5). These figures may in part reflect the fact that specific searches were made for these items; at the same time, however, the searches were made because of tendencies which had emerged during the grammatical searches. The next most common items, those with a frequency of between two and four, are shown in table 4.

Table 4: NOUN$_2$ items with frequency between 2 and 4

Tokens	
4	day, evening, free kick, job
3	book, doctor, horse, hotel, house, husband, interview, pass (in sport), place, shot (in sport), spider, story, wife
2	boy, city, clue, cousin, delivery (in sport), dog, father, hill, horseman, match, planet, residence, rush, time, village, voice, wave, year

The most common general meaning for N_2 is that of 'people', for which there were 165 tokens (45.2 percent of the total). Other notable areas of meaning are: a) sport, 30 tokens (e.g. 'a freak of a goal'); b) other forms of entertainment (especially music and computer games), 25 tokens (e.g. 'a Godfather of a track');

c) time reference, 15 tokens (e.g. 'a bastard of a year'); d) buildings or parts thereof, 15 tokens (e.g. 'that grave of a house'); e) fauna, 15 tokens (e.g. 'a beauty of a horse'). Since the N_2 slot was filled by a high proportion of 'people' nouns, I will now describe these in a little more detail.

A first point to make with regard to PEOPLE nouns relates to the sex of the people concerned: 109 are male, 43 female, and 13 unknown. By 'unknown' I mean that there is no indication in the text (for example, there may be reference to a 'doctor'), or that one would have to go far back in the text in order to discover it. The difference between the number of male and female referents is interesting, but could only be investigated properly by carrying out additional comparative studies, for example, by comparing the structure described in the present study with other ways of describing people. Quite a range of PEOPLE words are found in N_2 position, 31 different items referring to males, and 20 referring to females. As has already been mentioned, the most frequent items are the general nouns *man* (73) and *woman* (15). With regard to the difference in frequency between these two nouns, it is presumably of some relevance that in the corpus as a whole the word *man* is more than twice as common as the word *woman*.

Another aspect of PEOPLE nouns is the $NOUN_1$ phrases which are used to describe them. These may be seen in figure 2, which shows the individual items found with either male referents, female referents, or both (in CAPITALS). Some of the resulting phrases are:

(2) a fine-voiced giant of a singer (periodical: arts)
(3) a whirlwind of a boy (book: leisure)
(4) a stout bulwark of a woman (novel)
(5) his quiet mouse of a wife (book: arts)

$NOUN_1$s which describe $NOUN_2$ male referents:
ashpole, ass, bastard (3), beanpole (2), bear (8), BEAST, bird, blitz, brute (7), bull-dog, bullock, burke, devil (3), dynamo, eejit, fool (5), giant (17), gnome (2), grandmother (sic), hawk, hulk (3), humbug, hunk, jelly-heap, JEWEL, loner, maverick, midget, miracle, MONSTER, MOUNTAIN, MOUSE (2), peach, pixie, ponce, prat, prodigy, rock, rough diamond, runner bean, scallywag, scamp, scoundrel, scrag, SCRAP, shambles, shell, skeleton, slab (3), slob, sop, stick, stump, swine (2), thief, troll, villain, walrus, whippet, whirlwind, WISP, wizard, WRECK (2).

$NOUN_1$s which describe $NOUN_2$ female referents:
battleaxe (3), BEAST, bitch (8), bulwark, bundle, chit (5), dumpling, gem (2), genius, grande dame, hag, hippopotamus, JEWEL, lioness, marvel, MONSTER, MOUNTAIN (2), MOUSE (2), reed, SCRAP, shit, silly cow, slut, whore, WISP (2), WRECK.

Figure 2: $NOUN_1$s found with male and/or female referents in $NOUN_2$ position

Due to the generally low frequency of individual N_1s, not very much can be said about whether there is a preference for items to be used more with males or with females (leaving aside the fact that some of the N_1s are inherently male or female and will therefore normally be used with one or other sex, e.g. *bullock* and *bitch*). One thing that can be said is that there are a few items (*bear*, *brute* and *giant*) which clearly seem to prefer male referents.

Looking in particular at the very frequent N_2 *man*, we find that there is a greater tendency for it to be used with some $NOUN_1$s than with others. Specifically, *giant* and *bear* were found above all or exclusively with *man*, while *fool* was found only with PEOPLE words other than *man*. With regard to the $NOUN_2$ *woman* (15 tokens), no obvious lexical patterning was observed, with 13 different N_1s being used.

4.3 Modifiers

133 phrases (35 percent) include modifiers. These almost all appear before $NOUN_1$, with just five occurring before $NOUN_2$. An example of the latter is:

(6) a gem of an oak five-sided dressoir (periodical: arts)

In one case modifiers appear in both positions:

(7) a bald-headed bullock of a middle-aged man (novel)

Although the majority of modifiers appear before N_1, they only refer to this noun in about 30 cases. This usually happens when the modifiers are acting as intensifiers, but in a few cases they have a more descriptive function. Examples are:

(8) an absolute cracker of a video (periodical: world affairs)
(9) a real beauty of a goal (TV news script)
(10) this architectural gem of a palace (book: social science).

The majority of modifiers in pre-$NOUN_1$ position (about 70) qualify the second noun. Examples are:

(11) *The Fortune* is a tiny gem of a theatre (book: arts)
(12) who at 75 is a stern rock of a man (periodical: social science)
(13) that perspiring fool of a doctor (novel)

In examples (11) to (13), it is as if there were two modifers. In (11), for example, we are being told that *The Fortune* is both 'a tiny theatre' and 'a gem of a theatre'.

Sometimes the modifier and N_1 are similar in meaning, with the modifier serving to emphasize the noun:

(14) Sergeant Bragg, a great bear of a man (novel)
(15) a huge hippotamus of a woman was filling the doorway (book: social science)

In about 30 cases, the modifiers are best seen as qualifying the whole of the phrase that follows ($NOUN_1$ + $NOUN_2$), usually with an intensifier (examples (16) and (17)) but sometimes with a descriptive adjective (especially when N_2 is very general in nature) as in (18):

(16) It was an absolute nightmare of a day (newspaper)
(17) and a right swine of a job that was (novel)
(18) And that thin stump of a thing, that must surely be the pump (novel)

A final point regarding modifiers is that they occasionally form quite long strings:

(19) a big, happy-go-lucky but ambitious bear of a man (novel)
(20) a delightfully simple, wonderfully complex, utterly absorbing and thoroughly entertaining gem of a piece (miscellaneous printed text: arts)

4.4 The beginning of the phrases: *a* nightmare of a trip

Various determiners are found at the beginning of the phrases, though *a* and *an* are by far the most frequent items. Relative frequencies can be seen in table 5. The figure of 241 for *a*/*an* includes ten items in which there is also a predeterminer (*such* or *what*), and in these cases it might be more accurate to say that the phrases begin with *such a* or *what a*:

(21) But few people took seriously the idea of such a hulk of a man going without food (newspaper)

Table 5: Phrase-initial determiners

Item	*a/an*	*one*	*the*	*that*	*this*	poss. adj	poss. noun	*some* sing.	ZERO sing.	*some* pl.	ZERO pl.
Tokens	241	2	17	53	40	7	4	5	8	2	1

Examples with the less frequent phrase-initial elements are:

(22) He was without doubt *one* peach of a downhill racer (book: arts)
(23) Every stroke was applied with precision by *the* bitch of an SS officer (novel)
(24) But it's *her* bitch of a mother (novel)

(25) Aldridge, Malkin from *Irons'* peach of a pass, and then Aldridge again (newspaper)
(26) In her desperate condition, stuck on *some* rock of a planet (novel)
(27) It was near Christmas, that's all I know. Bitch of a night. It was raining (novel)
(28) it also has jewel*s* of villages like West Burton and Askrigg (book: leisure)
(29) THERE are *some* real gem*s* of bargains to be had in the old jewellery market (newspaper)

The use of plural nouns is quite unusual in this frame. There are a total of four such phrases in the database, one with no determiner (28), one beginning with *some of the*, and two with *some* (e.g. 29). In three cases, the second noun is *gems*, and in the other one it is *jewels*. We could tentatively conclude that this particular subset of NOUN_1 phrases (i.e. the 'precious stones' group) has expanded its usage to include plural phrases.

5. Text types

The *a* NOUN_1 *of a* NOUN_2 frame was found in many different text types, but was more heavily concentrated in one, that of 'imaginative literature', which effectively means published novels (together with the occasional piece of unpublished narrative writing). Imaginative literature makes up about 16 percent of the corpus so, of the 380 phrases in the database, natural distribution would predict that 60-61 items should appear from this sector. In actual fact, the figure is three times as high, 184 (48.4 percent).

A few points are worth making with regard to the relationship between text type and lexis. Firstly, a high proportion of NOUN_1 phrases found once only in the corpus were found in imaginative literature (54.4 percent), thus suggesting that such writing is more 'individual'. The following is an example:

(30) Mrs Burger, a little dumpling of a woman in linen trousers and a smock, like a Chinese peasant, was listening and smiling (novel)

Secondly, phrases which are structurally more complex, having more than one pre-modifier (25 in all), were found exclusively in written texts. Still with pre-modifiers, it was found that the more unusual lexis was typically found in descriptions of pop music, several from the same publication. Examples are:

(31) a throbbing, sexquake of a song (periodical: arts)
(32) a forehead-furrowingly serious frown of a song (periodical: arts)
(33) a curmudgeonly diamond of a song (periodical: arts)
(34) a dragging arseflap of a song (periodical: arts)
(35) a virile, pump action beast of a rock song (periodical: arts)

By contrast, the phrases found in the spoken sections of the corpus, 37 in all, were structurally simple. One is incomplete (no second noun), none of them have modifiers before N_2, and the only modifiers before N_1 are the semantically light word *little* (in 'a little bitch'), and four uses of the intensifier *real*. The latter were all in television sports news scripts (e.g. 'a real beauty of a goal').

Finally, a few of the repeated nouns in N_1 position tend to be found (or not found) in certain text types. *A cracker of a* (eight tokens) was found especially in newspapers and magazines, and was not found in novels. *A gem of a* (25 tokens) was found just in written language but only in texts other than novels. *A fool of a* (nine tokens) was found mainly in novels (eight tokens), and *a slab of a* (five tokens) and *a chit of a* (five tokens) were found only in novels.

6. The phrases within the text

In this section I deal with three aspects of communication: firstly, the status of the '*a* NOUN_1 *of a* NOUN_2' phrases from the point of view of the textual given and the textual new; secondly, the type of information being conveyed by the N_1 phrase, especially whether it is descriptive or evaluative; and thirdly, the syntactic position of the phrases.

6.1 The textual new: the whole phrase or part of the phrase

Broadly speaking, '*a* NOUN_1 *of a* NOUN_2' phrases may be divided into two categories: those in which the whole phrase is of importance from the point of view of information content, and those in which it is the NOUN_1 phrase which is the focus of attention. Many phrases of both types were found in the corpus. Examples of the first type are:

(36) Time and again Lear's long, witty, and entrancing letters were responded to, if at all, by what Lear called 'a nasty abortion of a note' (book: world affairs)
(37) The door-knocker was stiff and my nervous hand rapped too loudly. A great battleaxe of a woman opened the door a little. 'Yes?' (novel)
(38) He turned to Emily and kissed her hand and then took Hari in his arms in a bear hug of an embrace (novel)
(39) In Botha, they have a prolific goal-kicker, drop-goal expert and an elusive wisp of a fly-half (newspaper)

In phrases of the second type, the N_1 phrase provides new information and the N_2 merely reiterates a meaning which is to be found quite close by in the text. Most commonly, there is some sort of hyponymic relationship between the N_2 and a previously mentioned noun. The following are two examples, with the hypernyms *job* and *complaint*. Here, and in subsequent examples, the N_2 referent is in italics.

(40) Then later he remustered as a *stoker*, and a right swine of a job that was (novel)
(41) *Shingles*. I don't think it will be too bad a dose, but it's a beast of a complaint (novel)

More frequently, N_2 refers back to a proper noun or title of some sort:

(42) *Reckless Rufus* is a cracker of a game that no self-respecting C64 owner should be without (periodical: leisure)
(43) The early single *Like A Daydream*, their finest pop moment, was a monster of a song tonight (newspaper)
(44) *That* [the A59] is a pig of a road (radio phone-in)

In N_2 position there is sometimes a 'general' noun (e.g. *man*, *place*), which may also be considered a hypernym of sorts:

(45) We also found *Karl Bundt* […]. This bespectacled bear of a man had been born and lived here all his life (book: leisure)
(46) Only by the merest chance would a traveller through space happen upon *the planet Earth*; a tiny jewel of a place, orbiting an unexceptional star in an ordinary galaxy (book: applied science)

Other types of reiteration are also found, for example lexical repetition (47) and synonymy (48):

(47) but you know *the last year* has been a bastard of a year for us in terms of sickness (spoken: professional meeting)
(48) A *doctor* came and examined me; I moaned throughout the inspection. My wound was probed and bandaged, and then the fool of a medico bled me (novel)

6.2 General communicative functions of N1 phrases: evaluation and description

In all '*a* N_1 *of a* N_2' phrases, the first part of the phrase in some way modifies $NOUN_2$. Very often, modification involves 'evaluation', judged to be present in at least 80 percent of items in the database. This undoubtedly represents the order of magnitude, though 'evaluating' evaluation is at times a subjective matter, and in 17 cases I have not been able to make a definitive judgement as to whether or not an evaluative element is present. Example (49) is a case in hand.

(49) A square-jawed bull-dog of a man, he is Principal of the University of Rhodesia (unpublished text: social sciences)

Here, I wonder whether "bull-dog" is purely descriptive, appearing as it does alongside "square-jawed", or whether for some readers it would have an evaluative connotation, presumably positive since the wider context is favourable to the person being described.

Examples (50) to (55) may be considered as more or less pure evaluation, the first three positive and the second three negative.

(50) from nothing United burst into life and went and took the lead […]. A real beauty of a goal from Les Robinson (TV news script)
(51) she felt it had been a dream of an evening (novel)
(52) and this piledriver of a book, cousin to Scarfe and Steadman, might just provide a kick-start (book: leisure)
(53) This allowed relatively easy access from Bodø, the start of Kungsleden (Ammarnäs) being a sod of a place to reach (book: leisure)
(54) And then Nasser Hussain was caught in the gully as a brute of a delivery flew from just short of a length to take the shoulder of his bat (newspaper)
(55) If she hadn't been an awkward, obstinate bitch of a teenager, it would all have been very different (novel)

It is not always clear, however, where one should draw the line and determine that a phrase is 'only evaluative'. In (53) it could be argued that "a sod of a" is slightly descriptive since in this context it means 'difficult'.

Most evaluative N_1s were always found with a specific polarity of evaluation, and sometimes it is difficult to imagine them being used otherwise. For example, *gem*, *cracker* and *peach* are all used positively, and *bugger* and *bitch* are used negatively. However, this is not always the case. For example, *beast* is usually used negatively but can be positive, as in example (56).

(56) *Never Too Late* is a virile, pump action beast of a rock song (periodical: arts)

Interestingly, a few words, and the NOUN_1 phrases in which they appear, seem to have both negative and positive features, even within the same textual realization. Examples are *scamp* and *scallywag* ('that scamp of a grandfather of mine', 'that scallywag of a son of yours').

In examples (57) to (62), the N_1 phrases are not only evaluative (the first three negative, the second three positive), but also descriptive in some way.

(57) He was a grey-haired dynamo of a man, bursting with ideas and good humour (book: world affairs)
(58) Seljalandsfoss is a ponytail of a waterfall throwing itself clear of the rim of its cliff (book: leisure)
(59) We had eaten a good dinner (among other things a golden bolster of an omelette bursting its seams with truffles) (book: leisure)

(60) She beheld Rose in the act of taking the hand of a middle-aged battleaxe of a woman (novel)
(61) How could he forget the intense little man in that crypt of a dining room? (novel)
(62) There are continuing violations of human rights against the East Timorese while that farce of an inquiry is being carried out (Hansard)

A minority of N_1 phrases (about 13 percent) are just descriptive in nature. In most cases physical description is involved, as in (63) and (64), though more abstract description is also found (65).

(63) and mounted on a colossus of a horse very nearly the same colour as her hair (novel)
(64) I once lived next door to a giant of a man with feet like Yeti slippers (periodical: leisure)
(65) The sun dew is a freak of a plant (book: leisure)

A final point worth mentioning is that a few N_1 phrases may sometimes be considered to be intensifiers. The phrases in question are *a devil of a*, *a dickens of a*, *a humdinger of a*, and *a whale of a*. These may all be evaluative, as in examples (66) and (67).

(66) they are a devil of a mascot to parade about (newspaper)
(67) 'Oh, the sky is bright ...' - a mellow opening to a whale of a song (periodical: arts)

However, when N_2 is already evaluative, then its role could be considered as that of intensifying:

(68) But if you crossed to the wrong post there was a devil of an argument (TV news script)
(69) the musical that's a whale of a hit (TV news script)

An additional communicative role is that of wordplay, especially in the form of punning, present in twelve of the phrases in the database. The first nouns involved are *cracker*, *devil*, *dickens*, *gem* and *whale*. In all cases there is a sort of pun in which the writer (they are all in written texts) has elicited both the meaning of the NOUN_1 phrase (e.g. *a devil of a*) and an individual meaning of N_1 (e.g. *devil*). I will give three examples. In the first, there is an overlapping of the intensifying phrase *a devil of a* and the core meaning of the noun *devil*.

(70) CAN anyone help solve the devil of a mystery over a Middlesbrough landmark? A sandstone bridge at the meeting of Newham and Marton West becks in Acklam has been known as Devil's Bridge for 200 years but no one knows why (newspaper)

The second example involves the phrase *a whale of a*, which overlaps with the usual single-word meaning of *whale.*

(71) And that was certainly what the audience seemed to think of 'Moby Dick' at the curtain call, with a standing ovation and some joining in as the cast took a bow, it certainly is a whale of a tale (TV news script)

The third comes at the end of an advertisement for Christmas celebrations. In this example there is an overlapping of two phrases, *a cracker of a* and *Christmas cracker*.

(72) Have a Cracker of a Christmas at MAHON'S (miscellanea: leisure)

Most of the wordplay examples appear as succinct phrases at the beginning or end of news articles or advertisements. They all involve N_1 phrases which were found a number of times in the corpus, and which would presumably be recognized as familiar phrases by readers, thus guaranteeing the immediacy and success of the puns.

In addition to punning, there were also a few examples of other forms of wordplay, for example the rhyme in (71), and the juxtaposition of two unlikely nouns, as in (73):

(73) a great big boney giraffe of a horse (newspaper)

6.3 Syntax

Of the 380 phrases investigated, 329 are found in 'sentences', that is in complete grammatical structures with at least one finite verb. Their position within the clause is very variable, and they are found in all typical noun positions. Relative frequencies may be seen in table 6. The 'apposition' column refers to corpus tokens which are in apposition to the syntactic position indicated. An example is the already quoted (49), 'A square-jawed bull-dog of a man, he is Principal of ...', where the '*a* N_1 *of a* N_2' phrase is in apposition to the subject. The structure indicated in row seven (N_2 in 'N *of* N' phrase) can be exemplified with the already quoted (60): 'She beheld Rose in the act of taking *the hand of a middle-aged battleaxe of a woman*'.

The table has been kept as simple as possible, with various phenomena being subsumed under the general syntactic position indicated. Notably: (i) sometimes the '*a* N_1 *of a* N_2' phrase fills only part of the syntactic slot indicated; (ii) the 'complement' row includes complements which follow grammatical *it* or *there*; (iii) in a few cases, the subject and direct object are of a non-finite verb, as in the already quoted (21): 'But few people took seriously the idea of such a hulk of a man going without food'.

Table 6: Clause position of '*a* NOUN$_1$ *of a* NOUN$_2$' phrases

Syntactic position	Normal	Apposition	Total	%
Complement	102	13	115	35
Subject	54	15	69	21
Direct object	51	12	63	19
Prepositional phrase	39	2	41	12.5
Part of verbal phrase	15	-	15	4.6
Agent of passive	9	3	12	3.7
N_2 in 'N *of* N' phrase	6	2	8	2.4
Indirect object	5	1	6	1.8

Of the remaining phrases (those which are not part of finite-verb sentences), almost half are easily likened to complements, although there is no verb present. Examples are:

(74) United won a penalty: a real chop of a tackle. But Magilton missed, or rather Muggleton saved (TV news script)
(75) The sun beat down. A scorcher of a day. Heavy with insects. The pit was thick with flies (novel)

Other uses found a number of times are exclamations, as in (76), and introductory headings of some sort, as in (77):

(76) 'That devil,' she said. 'That bleeding bastard of a devil!' (novel)
(77) Gem of a clue. A RAIDER left a tell-tale clue after robbing a jeweller's – his parole papers (newspaper)

7. Concluding remarks

The phrases described and exemplified in the present article constitute a precise grammatico-semantic frame, and one which is very simple to define. Its essential grammatical feature is the presence of two nouns joined by the pivotal *of a/an*; its defining semantic feature is the relationship between the two nouns, whereby the first describes the second. It is a very open frame, in that new items can be created when thought desirable (e.g. 'a ponytail of a waterfall'), but one in which certain phrases become a part of common usage (e.g. 'a gem of a ...', '... of a man') or indeed lexicalized ('a whale of a time'). The frame has long formed part of the syntactic description of English, but it has not been possible to describe it in much detail because of the paucity of authentic examples.

For corpus linguistics, this type of frame is problematic, and this for a number of reasons. Firstly, it is *a priori* meaning based, and has no fixed lexis other than the phrase *of a/an* (of which there are about 147,000 tokens in the BNC): it is not for nothing that the frame appears in the OED under the word *of*. Secondly, the various grammatical forms it may take are part of more general

syntactic frames, which are realizable in other ways, semantically speaking (e.g. 'a quarter of an hour', 'this fear of a new apocalypse'). Thirdly, the current investigation suggests that it is not a particularly frequent frame in modern British English, especially if one excludes the NOUN_1 phrase *a hell of a*. In short, it is a beast of a frame to investigate, and automatic tools need to go hand in hand in both flexible search strategies and a considerable amount of reading of concordance lines.

References

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *The Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Bolinger, D. (1972), *Degree Words*. The Hague: Mouton de Gruyter.

Coffey, S. (2003), 'Phonically derived substitutes of potentially offensive lexical items: a corpus-aided study', in: C. Nocera, G. Persico and R. Portale (eds.) *Rites of Passage: Rational/Irrational, Natural/Supernatural, Local/Global*. Proceedings of the 20th conference of the Associazione Italiana di Anglistica. Soveria Mannelli: Rubbettino. 415-427.

Coffey, S. (2008), 'Remarks on the frequency and phraseology of *a/an* in modern English', in: A. Martelli and V. Pulcini (eds.) *Investigating English with Corpora: Studies in Honour of Maria Teresa Prat*. Monza: Polimetrica. 121-135. (Also available on-line at www.polimetrica.com).

Francis, G., S. Hunston and E. Manning (1998), *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

Meyer, C. (1992), *Apposition in Contemporary English*. Cambridge: Cambridge University Press.

Oxford English Dictionary, on-line edition. Oxford: Oxford University Press.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction[1]

Magali Paquot and Yves Bestgen

Université catholique de Louvain

Abstract

Most studies that make use of keyword analysis rely on log-likelihood ratio or chi-square tests to extract words that are particularly characteristic of a corpus (e.g. Scott and Tribble 2006). These measures are computed on the basis of absolute frequencies and cannot account for the fact that "corpora are inherently variable internally" (Gries 2006: 110). To overcome this limitation, measures of dispersion are sometimes used in combination with keyness values (e.g. Rayson 2003; Oakes and Farrow 2007). Some scholars have also suggested using other statistical measures (e.g. Wilcoxon-Mann-Whitney test) but these techniques have not gained corpus linguists' favour (yet?). One possible explanation for this lack of enthusiasm is that statistical tests for keyword extraction have rarely been compared. In this article, we make use of the log-likelihood ratio, the t-test and the Wilcoxon-Mann-Whitney test in turn to compare the academic and the fiction sub-corpora of the British National Corpus *and extract words that are typical of academic discourse. We compare the three lists of academic keywords on a number of criteria (e.g. number of keywords extracted by each measure, percentage of keywords that are shared in the three lists, frequency and distribution of academic keywords in the two corpora) and explore the specificities of the three statistical measures. We also assess the advantages and disadvantages of these measures for the extraction of general academic words.*

1. Introduction

One of the questions that has attracted most interest in corpus linguistics so far is: "Which words are particularly characteristic of a corpus?" (Kilgarriff 2001: 99). The simplest (and also most frequent) formulation of the problem has been to extract the keywords of a specific corpus, i.e. "items of unusual frequency in comparison with a reference corpus of some suitable kind" (Scott and Tribble 2006: 55), by means of log-likelihood ratio (or less frequently chi-square) tests. This is the only procedure that is currently implemented in widely used corpus linguistic tools such as WordSmith Tools (Scott 2004) and Wmatrix (Rayson 2003). Frequency has thus often been the sole criterion for identifying distinctive words of a particular corpus. In the last few years, a number of studies have

emphasized the need for taking corpus variability into account in corpus studies (cf. Gries 2006). In a study of vocabulary differences in English language corpora representing seven different countries, Oakes and Farrow (2007: 91) write that "[i]n a study of this nature, it is important to consider only those words which are relatively evenly spread throughout the corpus". In these conditions, it is questionable whether log-likelihood ratio (or chi-square) tests are always the best statistical measures to identify words that are particularly characteristic of a corpus.

The primary objective of this paper is to compare the number and type of keywords extracted from a corpus of academic writing by means of the log-likelihood ratio and two other statistical tests, viz. the t-test and the Wilcoxon-Mann-Whitney test. These two tests were selected as they take frequency distribution across corpus sections into account. A second objective is to assess whether one of these measures is better suited to identifying English for General Academic Purposes (EGAP) words, i.e. a rather formal vocabulary common to a wide range of academic texts but not so common in non-academic texts. Academic words "most probably occur because they allow academic writers to do the things that academic writers do. That is, they allow writers to refer to others' work (*assume*, *establish*, *indicate*, *conclude*, *maintain*); and they allow writers to work with data in academic ways (*analyse*, *assess*, *concept*, *definition*, *establish*, *categories*, *seek*)" (Nation 2001: 18). EGAP words also play an important part in discourse organisation and cohesion (cf. Halliday and Hasan 1976: 274-292; Partington 1998: 89-106; Nation 2001: 210-216). Put differently, these words "provide a semantic-pragmatic skeleton for the text. They determine the status of the (more or less technically phrased) propositions that are laid down in it, and the relations between them" (Meyer 1997: 9).

The paper is organised as follows. In section 2, we describe the most common procedure used in corpus linguistics to compare corpora and extract keywords. We highlight its major drawback, i.e. the fact that it does not take corpus variability into account, and briefly discuss other techniques proposed to overcome this limitation. Section 3 identifies the distinctive characteristics of the three statistical tests that we compare in this study. Section 4 describes the corpora and the methodology used to extract keywords from a corpus of academic texts. Section 5 provides a comparison of the keywords extracted by the three statistical tests and discusses the results. The article ends by identifying avenues for future research and revisiting the definition of 'keyword' in light of the results obtained in this study.

2. Keyword analysis

Keyword analysis has been used in a variety of fields to extract distinctive words or keywords, e.g. business English words (Nelson 2000), words typically used by men and women with cancer in interviews and online cancer support groups (Seale et al. 2006) and terminological items typical of specific sub-disciplines of

English for Information Science and Technology (Curado Fuentes 2001). As emphasized by Scott and Tribble, "keyness is a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail. What the text 'boils down to' is its keyness, once we have steamed off the verbiage, the adornment, the blah blah blah" (2006: 55-56).

The procedure to identify keywords of a particular corpus involves five main stages (cf. Scott and Tribble 2006: 58-60):

1. Frequency-sorted wordlists are generated for a reference corpus and the research corpus.
2. A minimum frequency threshold is usually set at 2 or 3 occurrences in the research corpus. Thus, "for a word to be key, then it (a) must occur at least as frequently as the threshold level, and (b) be outstandingly frequent in terms of the reference corpus" (Scott and Tribble 2006: 59).
3. The two lists of word types and their frequencies are compared by means of a statistical test, usually the log-likelihood ratio.
4. Words in the research corpus that do not occur at least as frequently as the threshold and statistically significantly more than the same type in the reference corpus are filtered out.
5. The wordlist for the research corpus is reordered in terms of the keyness of each word type. Software tools usually list positive keywords, i.e. words that are statistically prominent in the research corpus, as well as negative keywords, i.e. words that have strikingly low frequency in the research corpus in comparison with the reference corpus.

This method has been applied on different types of research corpora, from the most homogeneous to the most heterogeneous in terms of text type and domain. Scott and Tribble (2006: 179-193) carries out a keyword analysis on a single text, viz. Samuel Beckett's *Texts for Nothing, 1*. Tribble (1998) extracts keywords from a relatively homogeneous corpus of 14 project proposals concerned with the restructuring of social and financial institutions, thus representing one particular text type. Nelson (2000) identifies business English keywords from a heterogeneous corpus of business writing (emails, reports, faxes, etc.) and speech (meeting and negotiation transcripts and phone calls). The more heterogeneous a corpus, however, the less evident it is exactly what a keyword reveals about the research corpus. Leech et al. (2001) make use of the *British National Corpus* (BNC)[2] and list the keywords of very broad categories such as writing vs. speech, imaginative vs. informative writing and conversational vs. task-oriented speech. They first show that words such as *you*, *I*, *'s*, *yeah*, *it*, and *got* belong to the most distinctive keywords of the spoken part of the BNC compared to its written part (Leech et al. 2001: 218). However, a comparison of conversational vs. task-oriented speech (e.g. lectures, political speeches, legal proceedings, trade union talks and sports commentaries) later reveals that these words are not distinctive of any kind of speech, only of conversational speech (Leech et al. 2001: 242; see also Lee 2001a).

The keyword extraction procedure described above relies on the conception of a corpus as one big text rather than as a collection of texts. Statistical measures such as the log-likelihood ratio are computed on the basis of absolute frequencies and cannot account for the fact that "corpora are inherently variable internally" (Gries 2006: 110). As a consequence, the procedure cannot distinguish between global vs. local keywords. Global keywords are dispersed more or less evenly through the corpus while local keywords appear repeatedly in some parts of the corpus only, a phenomenon to which Katz (1996: 19) has referred as "burstiness".[3] For example, in a keyword analysis of gay male vs. lesbian erotic narratives, Baker shows that *wuz* (used as a non-standard spelling of *was*) appears to be a keyword of gay male erotic narratives when in fact its use is restricted to one single text, "which suggests that this word is key because of a single author's use of a word in a specific case, rather than being something that indicates a general difference in language use" (Baker 2004: 350). Put differently, the keyword status of *wuz* is more a function of the sampling decision to include one particular narrative than evidence of the distinctiveness of the word in gay male erotic narratives (see also Oakes and Farrow 2007: 91).

Three types of solution have been proposed in recent studies to overcome this serious limitation of keyword analysis. Some authors have built keyword databases. A keyword database reveals how many texts or sections in a research corpus a word appears in as key. A frequency-sorted wordlist is generated for each text/section in the research corpus and compared to a reference corpus wordlist. A list of keywords for each text/section is thus drawn. All these keyword lists are then compared to build a keyword database, with the requirement that, to be included, each keyword must appear in a minimum number of texts/sections (e.g. Tribble 2000; Mudraya 2006; Nelson 2000; Scott and Tribble 2006). As underlined by Scott:

> [A] 'key keyword' is one which is 'key' in more than one of a number of related texts. The more texts it is 'key' in, the more 'key key' it is. This will depend a lot on the topic homogeneity of the corpus being investigated. In a corpus of City news texts, items like *bank*, *profit*, *companies* are key key-words, while *computer* will not be, though *computer* might be a key word in a few City news stories about IBM or Microsoft share dealings. (Scott 2004)

Other researchers have re-classified keywords according to a measure of dispersion (e.g. Oakes and Farrow 2007; Paquot 2007a). The simplest measure of this type is range, i.e. a measure of frequency in terms of the number of texts or sections of the research corpus a keyword appears in (cf. Rayson 2003: 93-104). This method is less restrictive than building a key keyword database as a word does not need to be a keyword in all the sections it appears in. A more sophisticated measure is Juilland's D, i.e. "a statistical coefficient of how evenly distributed a word is across successive sectors of the corpus" (Rayson 2003: 93-104).[4] This measure takes into account "not only the presence or absence of a

word in each subsection of the corpus, but the exact number of times it appears" (Oakes and Farrow 2007: 91).

These two approaches, however, also have their drawbacks. As illustrated in the following two quotes, they both rely on an additional arbitrary cut-off point:

> A KKW [key keyword] is one which is key in lots of texts, where "lots" is defined (subjectively) by the number of texts in the database. In our present case where there are nearly 4,000 texts, a KKW would be one which occurs in say, 5% or more of the texts. (Scott and Tribble 2006: 78)

> Whether we use range, D [a measure of dispersion], or U [a usage coefficient which combines dispersion and frequency], our cut-off point for discriminating between well and poorly dispersed words must be arbitrary. (Oakes and Farrow 2007: 92)

Results of a keyword analysis are thus largely dependent on a number of arbitrary cut-off points: the probability threshold under which log-likelihood ratio values are not significant, a minimum frequency cut-off point, a minimum number of texts in which a keyword appears (as keyword or not) and/or a minimum coefficient of dispersion.

Kilgarriff (2001) proposes to make use of the Wilcoxon-Mann-Whitney test[5] in keyword analysis as this test takes dispersion (or corpus variability) into account, thus obviating the need for an additional arbitrary cut-off threshold. In a replication study of Rayson et al.'s (1997) keyword analysis of male vs. female speech, Kilgarriff states that this test is less sensitive to high absolute frequencies than the chi-square (χ^2) test. However, this statement is based on an analysis of the 25 most prominent keywords in male vs. female conversation according to the chi-square and Wilcoxon-Mann-Whitney tests and the author does not go into more detail about the differences between the two statistical measures. To our knowledge, the Wilcoxon-Mann-Whitney test has only been used in Kilgarriff (2001) and its parametric equivalent, viz. the t-test, has never been used in keyword extraction. This most probably stems from the common assumption that parametric tests, viz. tests based on the assumption of normal distribution, "are invalid in most cases of statistical text analysis unless either enormous corpora are used, or the analysis is restricted to only the very most common words" (Dunning 1993: 71). Parametric tests have nevertheless been used in word count approaches in psychological research, viz. approaches which try to link dimensions of word use to personality, demographic markers and differences in mental and physical health (cf. Hogenraad 1990; Oxman et al. 1988; Pennebaker et al. 2003; Rude et al. 2004; Spence 1980; Spence et al. 1978).

The linguistic added value (if any) of statistical measures such as the t-test and the Wilcoxon-Mann-Whitney test would be better appreciated if the keywords extracted by these statistical measures were compared to results of the log-likelihood ratio. There is a need for more comparisons of statistical tests on

the same corpus data so as to highlight the major characteristics of each measure.[6] The main objective of this paper is thus to compare keywords extracted from an academic writing corpus by means of the log-likelihood ratio, the t-test and the Wilcoxon-Mann-Whitney test. The three tests are described in the next section.

3. Three statistical tests under scrutiny

3.1 Log-likelihood ratio

The log-likelihood ratio is probably the most commonly used statistical test in keyword analysis. It is calculated on the basis of a contingency table such as table 1, which reads as follows: for word *w* in corpus A and B, there are *a* occurrences of *w* in A (which contains *a* + *c* words) and *b* occurrences in B (which totals *b* + *d* words). The test compares the observed frequencies in the table with expected frequencies, i.e. frequencies that would be expected if the null hypothesis (H_0) were true (Dunning 1993). In corpus comparison, and more specifically, keyword analysis, the null hypothesis is that both corpora consist of "words drawn randomly from some larger population" (Kilgarriff 2001: 99). Under the null hypothesis, words have the same probability of occurrence in each corpus and the "differences observed between two corpora have arisen by chance due to inherent variability in the data" (McEnery et al. 2006: 55).

Table 1: A contingency table

	Corpus A	Corpus B	
w	*a*	*b*	*a* + *b*
Not *w*	*c*	*d*	*c* + *d*
	a + *c*	*b* + *d*	*a* + *b* + *c* + *d* = N

If the difference between the observed and expected frequencies of a specific word is large, its log-likelihood ratio probability value (also p value) is close to 0 and the null hypothesis of independence can be rejected.[7] Words with p values smaller than an arbitrarily pre-established level of significance close to 0 are regarded as positive or negative keywords, depending on whether they are more or less frequent in the corpus under study.

3.2 T-test

The t-test for independent samples is a comparison-of-means test: it looks at the difference between the means from two different groups and evaluates whether "any difference found between the two groups [is] within what is expected due to chance for any two means in a particular population" (Oakes 1998: 13). If the difference is higher than expected by chance, the two samples can safely be regarded as coming from two different populations. To determine whether the difference is statistically significant, "we must place this difference in a sampling

distribution and discover how far it is from the central point of that distribution" (Oakes 1998: 13). The standard error of difference, viz. a statistical index that measures the range of values that the difference between the means could take if the two groups came from the same population, is used as a yardstick for that comparison. It is computed on the basis of the variance and size of each group sample. The Welch-Satterthwaite procedure may be used to solve the problem of heterogeneity of variance (Howell 2007: 202-203). To apply a t-test on corpus-derived frequencies, corpora are divided into sections or individual texts and word counts are computed for each section/text. When corpus sections differ in size, t-tests should preferably be computed on relative rather than absolute frequencies.

The choice of the t-test in keyword analysis could be criticized on the basis that it is "only valid where the data is normally distributed, which is not in general the case for word counts" (Kilgarriff 2001: 104). According to Howell (2007: 637), however, those who argue in favour of using parametric tests comment that "the assumptions normally cited as being required of parametric tests are overly restrictive in practice and that the parametric tests are remarkably unaffected by violations of distribution assumptions" (see also Rietveld et al. 2004: 360). In addition, the assumption of normality of the sampling distribution does not apply to the distribution of raw scores (or word counts) in the two groups but to the distribution of their mean and the difference between them. The central limit theorem states that the sampling distribution of the means approaches normality as the number of observations increases. In practice, it is generally suggested that a number of observations of 25 to 30 is sufficiently large to produce a normal sampling distribution (Howell 2007: 177, 203-204). Word counts, however, may be characterized by such a markedly skewed distribution that a larger sample size is often necessary.

3.3 Wilcoxon-Mann-Whitney test

The Wilcoxon-Mann-Whitney (WMW) test is generally regarded as the non-parametric equivalent of the t-test for two independent samples. However, it tests a slightly different null hypothesis, i.e. the hypothesis that the two samples "were drawn at random from identical populations (not just populations with the same mean)" (Howell 2007: 649). The WMW test is computed on ranked scores rather than word frequencies. Applied on corpus data, the WMW test first puts the frequencies of a given word *w* in each corpus section/text of the two corpora in rank order, from lowest to highest rank. For example, table 2 (below) shows that the word *w* does not occur in section 3 of corpus B, which gets the lowest rank (rank 1). Sometimes frequencies across corpus sections are identical. The word *w* occurs once in one section of corpus A and in two sections of corpus B. This situation is described as tied scores. Tied scores are assigned mean ranks: these three sections are thus attributed the mean of ranks 2, 3 and 4 (rank 3). Similarly, sections 2 and 3 of corpus A are assigned the mean of ranks 6 and 7 (rank 6.5).

The WMW test takes as input the sum of ranks assigned to the smaller group of observations, viz. the corpus with fewer sections, or, if the two corpora

have the same number of sections, the smaller of the two sums (cf. Howell 2007: 649). In table 2, for example, the corpus with fewer sections is corpus B and the sum of its ranks is 12. If each group is less than or equal to 25 observations, this sum can be evaluated against a Wilcoxon distribution table, which attributes a p value to each possible rank sum according to the number of observations comprised in each group. For larger groups, a normal approximation is used, including a continuity correction and a correction for the presence of tied scores (Siegel and Castellan 1988: 128-137; Bergmann et al. 2000: 73).

Table 2: Ranking procedure of the WMW test

	Corpus A					Corpus B			
Corpus section	1	2	3	4	5	1	2	3	4
Frequency	10	5	5	1	100	2	1	0	1
Rank	8	6.5	6.5	3	9	5	3	1	3

This test has been criticized in corpus linguistics on the basis that "[i]gnoring the actual frequency of occurrence […] means discarding most of the evidence we have about the distribution of words" (Rayson 2003: 47). However, Kilgarriff (2001: 115-116) has argued that, unlike statistical measures such as the chi-square test, the WMW test does not have a bias towards high-frequency items.

4. Data and methodology

The corpora used (table 3) are two sub-parts of the BNC, viz. a 100 million word collection of samples of written and spoken language from a wide variety of sources, designed to represent British English from the later part of the twentieth century. We made use of Lee's genre classification scheme (cf. Lee 2001b) in order to compile a corpus of academic texts (BNC-ACAD) and another of literary texts (BNC-LIT). The academic corpus consists of 501 published academic texts from several disciplines (e.g. humanities, medicine, natural science, politics, law, and engineering) and amounts to about 15 million words.[8] The heterogeneity of disciplines is particularly well suited to the purposes of this study as its second objective is to extract words that would be useful to all members of the academic "discourse community" (Swales 1990: 29).

Table 3: Corpora used

	BNC-ACAD	BNC-LIT
Number of texts	501	432
Number of words	15,429,582	15,926,677
Lee's (2001b) 'genre' labels	academic prose representing different disciplines (humanities; medicine; natural sciences; politics, law and education, social and behavioural sciences; technology, computing and engineering)	prose, fiction, drama

The literary corpus is of similar size but only contains 432 texts. BNC-LIT is not used as a normative corpus since it does not provide a text norm or general language standard against which BNC-ACAD can be compared (cf. Rayson 2003: 41). Instead, it is used as a "strongly contrasting reference corpus" (Tribble 2001: 396) on the hypothesis that useful words for writers from different academic disciplines would be particularly under-represented in literary texts.

We made use of the SAS statistical package to compare BNC-ACAD to BNC-LIT and extract its keywords with the help of the log-likelihood ratio, the t-test and the WMW test.[9] The three resulting lists were then compared on a number of criteria. We first examined the effects of three minimum frequency thresholds on the number of keywords extracted by each statistical measure: (1) no minimum frequency of occurrence, (2) a minimum frequency of one occurrence per million words, and (3) a minimum frequency of ten occurrences per million words in at least one of the two corpora. As the texts comprised in the two corpora differ in size, we used relative frequencies for the t-test and the Wilcoxon-Mann-Whitney test. A word was considered as a keyword according to one of the three tests when the probability resulting from the test was below an arbitrarily selected p value, as is the usual practice in natural language processing and corpus studies (cf. Moore 2004). We heuristically selected three threshold values: 0.01, 0.001 and 0.000001, the last one being the default value in WordSmith Tools for keyword analysis. Scott (2004) recommends the use of such a low p value to increase the selectivity of the extraction procedure and reduce the number of selected keywords.

We then examined keywords that are extracted by one, two or three measures and identified their major distributional characteristics. The main results of this study are discussed in the following section.

5. Results

Table 4 gives the number of keywords extracted from BNC-ACAD by the log-likelihood ratio, the t-test and WMW test at p value < 0.01, 0.001 and 0.000001 when no minimum frequency threshold is used. It shows that the log-likelihood ratio extracts many more keywords than the t-test and the WMW test. In addition, the number of keywords extracted by the log-likelihood ratio for $p < 0.000001$ decreases less markedly than the number of distinctive words selected by the two other measures. 51.35 percent of the keywords extracted by the log-likelihood ratio at $p < 0.01$ are also extracted at $p < 0.000001$. By contrast, only 33.89 percent and 40.33 percent of the keywords extracted at $p < 0.01$ by the t-test and the WMW test respectively are also extracted at $p < 0.000001$.

A comparison of tables 4, 5 and 6 shows that whatever the settings used, the log-likelihood ratio always extracts the highest number of keywords while the t-test almost always extracts the smallest number of distinctive words of academic writing. Second, the number of keywords extracted by the log-likelihood ratio for $p < 0.001$ and 0.000001 is very close to the number of

keywords extracted at $p < 0.01$ when minimum frequency thresholds are used. In fact, there is almost no point in using a $p < 0.000001$ when the minimum frequency threshold is set at ten occurrences per million words: almost 96 percent of the keywords extracted at $p < 0.01$ by the log-likelihood ratio are also extracted at $p < 0.000001$ (cf. table 6). Our results thus show that, given enough data, the log-likelihood ratio attributes very extreme p values, that is, p values that are close to zero. These results can be paralleled with Kilgarriff's (2005: 268) general comment on null hypothesis testing by means of chi-square tests[10] that "[g]iven enough data, H_0 [the null hypothesis] is almost always rejected however arbitrary the data".

By contrast, results of the WMW test and, more particularly, the t-test differ widely according to the p value used. When a minimum frequency threshold of one occurrence per million words is used, only 35.43 percent of the keywords extracted by the t-test at $p < 0.01$ are also extracted at $p < 0.000001$ (cf. table 5).

Table 4: No minimum frequency

	Log-likelihood ratio		t-test		WMW	
0.01	26,387	100%	8,224	100%	10,031	100%
0.001	20,882	79.12%	5,486	66.71%	7,048	70.26%
0.000001	13,551	51.35%	2,771	33.89%	4,046	40.33%

Table 5: Minimum frequency of 1 per million words

	Log-likelihood ratio		t-test		WMW	
0.01	18,110	100%	7,821	100%	9,424	100%
0.001	17,176	94.84%	5,468	69.91%	7,025	74.54%
0.000001	13,551	74.82%	2,771	35.43%	4,046	42.93%

Table 6: Minimum frequency of ten per million words

	Log-likelihood ratio		t-test		WMW	
0.01	5,551	100%	4,231	100%	4,155	100%
0.001	5,478	98.68%	3,615	85.44%	3,864	93%
0.000001	5,324	95.91%	2,419	57.17%	3,125	75.2%

This paper also has the objective of assessing whether one of the three statistical measures is better suited to extracting words that would be useful for writers over a wide range of academic disciplines. For a keyword to be part of such an EGAP vocabulary, it should be relatively frequent in academic writing. In the remaining part of this paper, we therefore focus on results for the following settings: a minimum frequency threshold of ten occurrences per million words and a p value set at $p < 0.000001$.

An analysis of the 10,333 different types that appear ten times or more in at least one of the two corpora shows that, if the log-likelihood ratio is used, almost all of them are distinctive of academic or literary texts. As shown in table

7, more than half of the types are academic keywords (51.5 percent) and 40.2 percent are keywords of BNC-LIT. The t-test is much more selective and classifies 51.5 percent of the types as statistically non-significant while the WMW test occupies a middle ground with 33.8 percent. Another difference between the log-likelihood ratio and the two other tests is that the former classifies a larger proportion of types as keywords of BNC-ACAD while the t-test and the WMW test identify a higher proportion of keywords in BNC-LIT.

Table 7: Minimum frequency of ten per million words, $p < 0.000001$

Keywords in …	Log-likelihood ratio		t-test		WMW	
BNC-LIT	4,158	40.2%	2,588	25%	3,718	36%
-	851	8.2%	5,326	51.5%	3,490	33.8%
BNC-ACAD	5,324	51.5%	2,419	23.4%	3,125	30.2%
	10,333	100%	10,333	100%	10,333	100%

Figure 1 focuses on the 5,324, 2,419 and 3,125 keywords in BNC-ACAD extracted by the log-likelihood ratio, the t-test and the WMW test respectively. It shows that 2,262 keywords are extracted by the three statistical measures, representing 93.5 percent and 72.4 percent of the total number of keywords extracted by the t-test and the WMW test but only 42.5 percent of the keywords selected by the log-likelihood ratio. Figure 2 gives the first 200 shared keywords that have the lowest p values according to the three tests. They correspond rather well to our definition of EGAP words. A large proportion of these words are used to structure academic texts and express cause and effect (e.g. *arise*, *arises*, *arising*, *consequence*, *consequently*), purpose (e.g. *aim*, *aims*) comparison and contrast (e.g. *compare*, *compared*, *comparison*, *contrast*), concession (e.g. *albeit*, *although*), opposition (e.g. *contrary*, *conversely*). Others "have in common a focus on research, analysis and evaluation – those activities that characterize academic work" (Martin 1976: 92). They are used to describe steps of scientific research (e.g. *analyse*, *analysis*, *approach*, *classification*, *classified*, *conducted*), to refer to abstract ideas and processes (e.g. *ability*, *assumption*, *basis*, *case*, *category*, *combination*, *concept*), to refer to the ideas and findings of others (e.g. *argue*, *argument*, *claim*, *claimed*, *conclude*), and to evaluate them (e.g. *accept-able*, *accurate*, *accurately*, *adequate*, *certain*, *clear*, *correct*).

Other keywords of BNC-ACAD are extracted by one measure only. As shown in figure 1, the WMW test only extracts seven keywords on its own and the t-test only extracts keywords that are extracted by at least one other measure. By contrast, 2,049 keywords are only extracted by the log-likelihood ratio (cf. figure 3). Most of these keywords are topic-dependent or discipline-related words and do not qualify as EGAP vocabulary as they are restricted to one or two academic disciplines. For example, 79 percent of the total number of occurrences of the word *adjective* (310/390) appear in a single text entitled: "The meaning of syntax: a study in the adjectives of English". The word *antral* appears 182 times

in BNC-ACAD but only occurs in six medicine texts. Similarly, 99 percent of the occurrences of the word *appellants* appear in 26 texts classified by Lee (2001b) under the 'politics, law and education' category.

	t-test	Log-likelihood ratio	WMW
	---------- 2262 ----------		
	[93.5%]	[42.5%]	[72.4%]
	---------- 157 ----------		0
	[6.5%]	[3%]	
	0	---------- 856 ----------	
		[16%]	[27.4%]
	0	2049	7
		[38.5%]	[0.2%]
Total	2419 [100%]	5324 [100%]	3125 [100%]

Figure 1: Distribution of shared keywords in BNC-ACAD

ability, absence, acceptable, accepted, according, accordingly, account, accurate, accurately, achieve, achieved, achieving, acquire, act, action, active, actively, activities, activity, acts, actual, addition, additional, adequate, adequately, administration, adopt, adopted, advantages, advocated, affect, affected, affecting, affects, aim, aims, albeit, allow, allowing, allows, also, alternative, alternatively, although, amount, amounts, an, analyse, analysed, analysis, apparent, appear, appears, applied, applies, apply, applying, approach, appropriate, approximately, are, area, areas, argued, argument, arguments, arise, arises, arising, as, aspect, aspects, assess, associated, association, assume, assumes, assumption, assumptions, attempt, attempts, attitudes, attributed, availability, available, average, balance, based, basic, basis, be, become, becomes, behaviour, being, between, both, broader, broadly, by, can, cannot, case, cases, categories, category, central, centres, century, certain, change, changes, chapter, characteristic, characteristics, characterized, circumstances, cited, civil, claim, claimed, claims, classes, classification, classified, clear, clearly, combination, combined, commentators, commitment, common, commonly, community, compare, compared, comparing, comparison, complex, complexity, concept, concepts, concern, concerned, concerning, concerns, conclude, concluded, conclusion, conclusions, conditions, conducted, confined, conflict, conform, consequence, consequences, consequent, consequently, consider, considerable, considerably, considerations, considered, considering, consist, consistent, consistently, consists, constitute, constituted, constitutes, contain, contained, contains, contemporary, content, context, continues, continuing, continuity, continuous, contrary, contrast, contribute, contributed, contributions, control, controlled, conversely, correct, courts, create, created, creating, creation, crucial, cultural, culture, currently, data

Figure 2: First 200 shared keywords in BNC-ACAD (alphabetical)

abdominal, abnormality, abolitionists, abortion, absorption, abundance, abundant, abuse, acceleration, accession, accidents, accommodate, accommodation, accompanying, accumulator, accused, acid, activated, activating, activator, add, addresses, adhesion, adjective, adjectives, adjudication, admission, adorno, advertising, advice, adviser, affinity, Afghanistan, Africa, African, afro-Caribbean, aged, ageing, agent, ages, aggression, aggressive, agrees, albumin, alcohol, algorithms, alkaline, allegations, allies, allowance, allowances, Althusser, America, American, amino, amongst, amplitude, anaemia, anal, analyst, angles, Anglia, Anglo-Saxon, answers, antibiotics, antibodies, antibody, antigen, antigens, antislavery, antral, apical, appellants, appointment, appointments, appraisal, appropriation, approval, Aquitaine, arbitrator, archaic, archbishop, architecture, archive, artic, aristocracy, Aristotle, arithmetic, arousal, arrangement, array, art, artefacts, arterial, artery, artificial, artistic, artists, arts, assault

Figure 3: First 100 keywords extracted by the log-likelihood ratio only

Among the 2,049 keywords of academic writing exclusively extracted by the log-likelihood ratio, 67 are classified as keywords of fiction writing by the Wilcoxon-Mann-Whitney test (cf. figure 4). These contradictory results are typically produced for words that are more frequent in the academic corpus but which occur in a larger number of texts and are therefore more evenly distributed in the fiction corpus (cf. table 8 below).

angle, animals, bile, cabinet, care, chain, child, children, china, company, crust, curve, deaf, elderly, emptying, estate, family, fans, file, firm, French, hearing, hospital, insects, land, landscape, lanes, library, load, lone, middle, months, mouse, movies, muscle, newspapers, notice, oil, older, outer, owl, owner, parents, patient, people, plants, plates, police, pound, prey, prison, rats, school, showed, smoking, soil, strings, surface, tale, television, trust, unconscious, women, word, words, world, worms

Figure 4: Log-likelihood ratio vs. WMW test: 67 keywords with contradictory results

Table 8: Distribution of the words *animals*, *family* and *landscape* in BNC-ACAD and BNC-LIT

	BNC-ACAD		BNC-LIT	
	Freq.	N° of texts	Freq.	N° of texts
animals	1826	143 [28.5%]	787	232 [53.7%]
family	5904	339 [67.7%]	4620	399 [92.4%]
landscape	502	73 [14.6%]	335	158 [36.6%]

A closer examination of the WMW values for the other keywords extracted exclusively by the log-likelihood ratio reveals that 421 types get negative WMW values even though their p values are not significant at 0.000001. The WMW test attributes negative values to types that are considered to be more distinctive of the fiction corpus. More than 20 percent of the types classified by the log-likelihood ratio as exclusively academic keywords are thus categorized as more characteristic of the fiction corpus by the WMW test. The word *heroin*, for example, is classified as a keyword of BNC-ACAD by the log-likelihood ratio (p value close to 0) but is a keyword of BNC-LIT according to the WMW ($p = 0.0033$). Other examples include *kings*, *peasant*, *war*, *victory*, *episode*, *tissue*, *suffering* and *witchcraft*.

Finally, figure 1 also shows that there are keywords of BNC-ACAD that are extracted by two measures. 157 keywords are extracted by the t-test and the log-likelihood ratio but not by the Wilcoxon-Mann-Whitney and 856 keywords are extracted by the log-likelihood ratio and the Wilcoxon-Mann-Whitney but not by the t-test. Without entering into too many details, a few general tendencies seem to emerge. The keywords extracted by the log-likelihood ratio and the t-test (cf. figure 5) are often characterized by the following pattern of behaviour: they appear with very low frequencies in a large number of academic texts but are also frequent in a limited set of texts. They are not retrieved by the WMW test as the two phenomena cancel each other out when compared to a corpus in which these words appear in many texts with frequencies that are less extreme. In terms of rank, they get very high and very low ranks in BNC-ACAD and intermediary ranks in BNC-LIT. By contrast, they are extracted by the log-likelihood ratio as they are more frequent in BNC-ACAD than in BNC-LIT and by the t-test as texts with high frequencies increase the mean score.

The keywords extracted by the log-likelihood ratio and the Wilcoxon-Mann-Whitney appear in a very small set of academic texts but they have very high frequencies in these texts. They are extracted by the log-likelihood ratio as they are more frequent in BNC-ACAD while the WMW test selects them as they get the highest ranks. They are characterized by a very uneven distribution in the corpus, which explains why they are not retrieved by the t-test, and are often topic-dependent or discipline-specific (cf. figure 6).

In summary, the only quantitative information needed to perform a log-likelihood ratio is the total number of occurrences of a word in two corpora as it does not take account of word distribution in corpus sections or texts. As a result, more keywords are extracted by the log-likelihood ratio, including discipline or topic-dependent words that are very unevenly distributed and appear repeatedly in some parts of the corpus only. These findings are in line with Gries's comment that "null hypothesis testing by means of chi-square tests does not appear to be a truly fruitful strategy for the word-frequency comparison of corpora" (Gries 2005: 281).

acted, addressed, age, appealed, articulate, attached, attendance, avoided, believes, borne, bound, brings, broad, caused, changing, character, chosen, clinical, colleagues, comment, commissions, completed, composed, concentration, confirmed, consciously, consisted, convenience, convert, convictions, cost, county, court, criminal, crisis, deal, decide, defended, department, depended, depicted, depth, details, detect, diseases, dismissed, duty, eastern, efforts, elaborate, English, enter, estates, experienced, failed, families, figure, first, follow, foreign, forming, freedom, greatest, handling, health, helpful, helps, high, human, imaginative, imposing, inevitable, intention, intentions, interviews, item, leaders, learning, least, length, liberties, locate, medical, mental, mentioned, mere, middle-class, million, movement, mutual, negotiated, neighbouring, November, occasions, opinions, ordinary, organised, overcome, party, person, phrase, planning, politics, premature, private, professional, prolonged, properly, prospective, prosperity, protect, qualities, questions, raise, random, rapidly, reason, receipt, recognised, recognising, reflected, regularly, respond, revealed, rightly, risk, ruled, scrutiny, sense, sensitive, sentence, September, serious, several, severity, sexes, shared, some, staff, striking, suffer, supreme, surplus, surrounding, targets, teaching, therapeutic, trade, training, union, unless, unlike, unnecessary, whom, work, works, writing

Figure 5: 157 keywords extracted by the log-likelihood ratio and t-test

abilities, abnormalities, abolition, abstraction, ac, academics, accountability, accounting, accumulation, acids, activation, activists, ad, additions, administrative, administrators, admissions, adoption, adult, advance, advent, advisory, advocates, aesthetic, affective, affidavit, agency, agents, aggregation, aid, aided, aids, al, algorithm, alienation, alliance, alliances, allied, alteration, alterations, amendments, amplification, analysts, analytic, analytical, Anderson, annum, anthropological, anthropologists, anthropology, apparatus, appellant, applicant, applicants, appoint, appointed, approximation, arbitration, arena, arguably, articulation, Asia, Asian, assemblies, assessments, assimilation, assisted, attach, attained, attainment, attorney-general, attribute, audit, authorised, awarded, axis, bacterial, basal, base, baseline, bases, bears, beneficial, beneficiaries, bias, biased, bilateral, binary, binding, biology, borough, boroughs, boundary, bourgeoisie, breaches, bureaucracy, bureaucratic, calcium, candidate, candidates

Figure 6: First 100 keywords extracted by the log-likelihood ratio and the WMW test

Unlike the log-likelihood ratio, the t-test and the WMW test are performed on word frequencies per text/section in corpora and thus take word distribution into account. However, these two measures differ in the way they deal with word distribution. The WMW test is based on a substitution of word frequencies by rank scores and does not take account of proper word frequencies. By contrast,

the t-test takes account of word frequencies in corpus sections to compute a mean and is therefore sensitive to outliers, viz. extreme values. For example, table 9 shows that the word *annum* appears in 56 academic texts and three fiction texts. The table should read as follows: in BNC-ACAD, the word *annum* appears once in 26 texts, twice in 17 texts and 49 times in a single text. It is classified as a keyword of BNC-ACAD by the WMW test as it gets the highest ranks with text frequencies of 49, 13, 10, etc. It is, however, not a keyword according to the t-test. Its mean in BNC-ACAD is higher than in BNC-LIT but the outlier (score 49) significantly increases the variance and the standard error of difference, which results in a non-significant t-test. A typical example of a keyword according to the t-test is *asserts* (cf. table 10) which is characterized by a relatively even distribution of frequency values that are included within a limited set of scores (from 1 to 6) and no outliers.

Table 9: Distribution of the word *annum* in BNC-ACAD and BNC-LIT

Frequency	BNC-ACAD texts	BNC-LIT texts
1	26	3
2	17	0
3	3	0
4	2	0
6	3	0
8	1	0
9	1	0
10	1	0
13	1	0
49	1	0
Total	56	3

Table 10: Distribution of the word *asserts* in BNC-ACAD and BNC-LIT

Frequency	BNC-ACAD texts	BNC-LIT texts
1	64	3
2	23	0
3	3	0
4	4	0
5	2	0
6	2	0
Total	98	3

These characteristics of the t-test make it a reasonably good test to identify EGAP vocabulary, especially when compared to the log-likelihood ratio which also gives prominence to discipline or topic-dependent words. We have argued that

the 2,262 keywords extracted by the log-likelihood ratio, the t-test and the WMW test are good potential candidates for inclusion in an EGAP vocabulary as they are lexical means necessary to do the things that academic writers do, e.g. stating a topic, hypothesizing, contrasting, exemplifying, explaining, evaluating, etc. The results of the t-test can be regarded as the best approximate to this common core of EGAP words as it only extracts 157 additional keywords that are shared by the log-likelihood ratio but not by the WMW test.

6. Future work

In the future, we would like to follow three avenues of research. The preliminary results presented in this paper have shown that there are often several word forms of the same lemma that are selected as keywords in academic writing (e.g. *arise*, *arisen*, *arises*, *arising*, *arose* or *determine*, *determined*, *determines*, *determining*). As rightly pointed out by Sinclair (1991), lemmas are an abstraction and only using lemmas would amount to losing important information as each word form has its own individual patterning. It may, however, be useful to replicate our study and use lemmas as units of analysis to give a general overview of academic vocabulary (see also Granger and Paquot in press).

As shown in section 5, the log-likelihood ratio, the t-test and WMW test do not select the same number of keywords, the log-likelihood ratio picking up many more keywords than the two other tests. This is a direct consequence of the use of an arbitrarily selected threshold probability value under which words are significant keywords. It would be interesting to apply the procedure recommended by Manning and Schütze (1999) and Evert (2004, 2008) for the study of collocations to keyword extraction: rank the words in decreasing order of p value and select the *n* highest ranking candidates, which are also referred to as an *n*-best list. The advantage of such a procedure is that "it allows for a 'fair' comparison of different measures because exactly the same number of candidates are evaluated from each ranking" (Evert 2004: 139).

Finally, we have relied on a fairly intuitive definition of the concept of 'EGAP vocabulary'. The purpose of this article has been to compare three statistical tests and assess their advantages and disadvantages for the extraction of general academic words. However, we did not rely on a gold standard of academic words (cf. Evert 2004) as it is our intention to build such a list (see also Paquot 2007b). A way out of this problem might be to have English for Academic Purposes (EAP) instructors evaluate the pedagogical importance of the keywords extracted by each test.

7. Conclusion

Our study has shown that the selection of a statistical test strongly influences the type of results obtained in keyword extraction. The log-likelihood ratio, the t-test and the Wilcoxon-Mann-Whitney test answer quite differently the question of

"[w]hich words are particularly characteristic of a corpus" (Kilgarriff 2001: 99). When the log-likelihood ratio is used, the sole criterion for keyword extraction is a higher absolute frequency of occurrence in the corpus under study than in a reference corpus. The t-test and the WMW test, however, also take account of the number of corpus sections in which a word occurs. The t-test is also sensitive to evenness of distribution. The log-likelihood ratio has won the favour of many corpus linguists, most probably because it is one of the few tests that is usually implemented in corpus linguistic (CL) tools. Currently, if we want to make use of other statistical tests, we need to leave most CL tools aside and make use of statistical software packages such as R, SAS or SPSS. As these tests are not particularly difficult to implement, it is to be hoped that they will make their appearance in CL tools in the near future.

Instead of being conditioned by available CL tools, the selection of a statistical measure should be dependent on the research question. For our purposes, viz. identifying EGAP words, the t-test proves to be better suited as it takes into account the distribution of words across the different texts and extracts relatively few topic-dependent or discipline-specific words compared to the log-likelihood ratio and the Wilcoxon-Mann-Whitney test. In fact, the t-test seems to satisfy Baker's (2004: 351) demand for "a way that combines the strength of key keywords with those of keywords but is neither too general or exaggerates the importance of a word based on the eccentricities of individual files".

Our study also points to the need for refining the concept of 'keyword'. If one single text is analysed, keywords are best defined as 'items of unusual frequency in comparison with a reference corpus of some suitable kind' and the log-likelihood ratio is a good test to extract them. However, as soon as corpora are analysed, the concept of keyword would be much more useful if it also relied on word distribution. If the keyword extraction procedure takes account of word distribution across corpus sections it will be more selective and avoid topic-dependency or idiosyncrasies of individual texts. To conclude, we would therefore like to propose a modest amendment to the definition of 'keyword':

> Keywords of a specific corpus are lexical items that are evenly distributed across its component parts (corpus sections or texts) and display a higher frequency and a wider range than in a reference corpus of some kind.

As shown in this study, a possible way of extracting keywords following this definition is to make use of the t-test.

Notes

1 We gratefully acknowledge the financial support of the Belgian National Fund for Scientific Research (F.N.R.S) and the *Communauté française de Belgique*. We also thank Professor Sylviane Granger for helpful comments on a previous version of this paper.

2 The *British National Corpus* (BNC) contains approximately 100 million words which reflect a wide variety of text types, genres and registers. The written component totals 90 percent of the corpus and includes samples of academic books, newspaper articles, popular fiction, letters, university essays and many other kinds of text. The spoken component represents ten percent of the whole corpus and consists of monologues and dialogues in various contexts, e.g. business, leisure and education. For more information on the BNC, see http://www.natcorp.ox.ac.uk.

3 Katz (1996: 19) distinguishes between "document-level burstiness", i.e. "multiple occurrences of a content word or phrase in a single-text document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all"; and "within-document burstiness" or "burstiness proper", i.e. "close proximity of all or some individual instances of a content word or phrase within a document exhibiting multiple occurrence".

4 Measures of dispersion do not necessarily need to be used in addition to keyness values. Zhang et al. (2004), for example, have used a measure of dispersion to define a core lexicon on the basis that "if a word is commonly used in a language, it will appear in different parts of the corpus. And if the word is used commonly enough, it will be well-distributed". The reader is referred to Oakes (1998: 189-192) and Gries (2008) for more information on measures of dispersion.

5 Following Bergmann, Ludbrook and Spooren (2000), we use the term 'Wilcoxon-Mann-Whitney test' to refer to two equivalent tests that were developed independently, viz. the Wilcoxon rank sum test and the Mann-Whitney U test.

6 Chujo and Utiyama (2006) and Chujo et al. (2007) compare nine statistical tests for keyword extraction but the selected tests are similar in that they are applied on absolute frequencies of occurrence. None of the tests under study takes corpus variability into account.

7 The log-likelihood ratio has a distribution similar to that of the chi-square. A statistical table for the distribution of the chi-square test can thus be used to find the log-likelihood ratio probability value.

8 The fact that some texts have been truncated in the *British National Corpus* is not ideal but the BNC academic sub-corpus is the largest corpus of academic texts from different disciplines that we could find.

9 We made use of the normal approximation to determine p values of the WMW test.

10 When applied to large sample sizes, log-likelihood ratio tests and Pearson chi-square tests are equivalent tests (Dunning 1993; Howell 2007: 152)

References

Baker, P. (2004), 'Querying keywords: questions of difference, frequency and sense in keyword analysis', *Journal of English Linguistics*, 32(4): 346-359.

Bergmann, R., J. Ludbrook and W. Spooren (2000), 'Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages', *The American Statistician*, 54(1): 72-77.

Chujo, K. and M. Utiyama (2006), 'Selecting level-specific specialized vocabulary using statistical measures', *System*, 34: 255-269.

Chujo, K., M. Utiyama and T. Nakamura (2007), 'Extracting level-specific science and technology vocabulary from the Corpus of Professional English (CPE)', in: M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds.) *Corpus Linguistics Proceedings 2007*. Available online from: http://www.corpus.bham.ac.uk/corplingproceedings07/.

Curado Fuentes, A. (2001), 'Lexical behaviour in academic and technical corpora: implications for ESP development', *Language Learning and Technology*, 5(3): 106-129.

Dunning, T. (1993), 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics*, 19(1): 61-74.

Evert, S. (2004), *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Available from: http://www.collocations.de/phd.html.

Evert, S. (2008), 'Corpora and collocations', in: A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. Article 58.

Granger, S. and M. Paquot (in press), 'Lexical verbs in academic discourse: a corpus-driven study of learner use', in: M. Charles, D. Pecorari and S. Hunston (eds.) *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum.

Gries, S. (2005), 'Discussion note: null hypothesis significance testing of word frequencies: a follow-up on Kilgarriff', *Corpus Linguistics and Linguistic Theory*, 1(2): 277-294.

Gries, S. (2006), 'Exploring variability within and between corpora: some methodological considerations', *Corpora*, 1(2): 109-151.

Gries, S. (2008), 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics*, 13(4): 403-437.

Halliday, M. and R. Hasan (1976), *Cohesion in English*. London: Longman.

Hogenraad, R. (1990), *A Little Organon of Content Analysis: From the Psychological Analysis of Discourse to the Analysis of Psychological Discourse*. Habilitation thesis in Psychology, Université catholique de Louvain.

Howell, D. (2007), *Statistical Methods for Psychology*. Belmont: Thomson-Wadsworth.

Katz, S. (1996), 'Distribution of common words and phrases in text and language modelling', *Natural Language Engineering*, 2(1): 15-59.

Kilgarriff, A. (2001), 'Comparing corpora', *International Journal of Corpus Linguistics*, 6(1): 97-133.

Kilgarriff, A. (2005), 'Language is never, ever, ever random', *Corpus Linguistics and Linguistic Theory*, 1(2): 263-275.

Lee, D. (2001a), 'Defining core vocabulary and tracking its distribution across spoken and written genres: evidence of a gradience of variation from the British National Corpus', *Journal of English Linguistics*, 29(3): 250-278.

Lee, D. (2001b), 'Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle', *Language Learning and Technology*, 5(3): 37-72. Available online from: http://llt.msu.edu /vol5num3/lee/default.html.

Leech, G., P. Rayson and A. Wilson (2001), *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* London: Longman.

Manning, C. and H. Schütze (1999), *Foundations of Statistical Natural Language Processing*. Cambridge and Massachusetts: MIT press.

Martin, A. (1976), 'Teaching academic vocabulary to foreign graduate students', *TESOL Quarterly*, 10(1): 91-97.

McEnery, A., R. Xiao and Y. Tono (2006), *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

Meyer, P.G. (1997), *Coming to Know: Studies in the Lexical Semantics and Pragmatics of Academic English*. Tübingen: Gunter Narr.

Moore, R.C. (2004), 'On log-likelihood ratios and the significance of rare events', in: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. 333-340.

Mudraya, O. (2006), 'Engineering English: a lexical frequency instructional model', *English for Specific Purposes*, 25(2): 235-256.

Nation, P. (2001), *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nelson, M. (2000), *A Corpus-Based Study of Business English and Business English Teaching Materials.* Unpublished PhD thesis, University of Manchester.

Oakes, M. (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oakes, M. and M. Farrow (2007), 'Use of the Chi-Squared Test to examine vocabulary differences in English language corpora representing seven different countries', *Literary and Linguistic Computing*, 22(1): 85-99.

Oxman, T.E., S.D. Rosenberg, P.P. Schnurr and G.J. Tucker (1988), 'Diagnostic classification through content analysis of patients' speech', *American Journal of Psychiatry*, 145(4): 464-468.

Paquot, M. (2007a), *EAP Vocabulary in Native English and EFL Learner Writing: From Extraction to Analysis. A Phraseology-Oriented Approach*. Unpublished PhD thesis, Université catholique de Louvain.

Paquot, M. (2007b), 'Towards a productively-oriented academic word list', in: J. Walinski, K. Kredens and S. Gozdz-Roszkowski (eds.) *Corpora and ICT in Language Studies. PALC 2005*. Frankfurt am Main: Peter Lang. 127-140.

Partington, A. (1998), *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam and Philadelphia: John Benjamins.

Pennebaker, J.W., M.R. Mehl and K.G. Niederhoffer (2003), 'Psychological aspects of natural language use: our words, our selves', *Annual Review of Psychology*, 54: 547-577.

Rayson, P. (2003), *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. Unpublished PhD thesis, Lancaster University.

Rayson, P., G. Leech and M. Hodges (1997), 'Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus', *International Journal of Corpus Linguistics*, 2(1): 133-152.

Rietveld, T., R. van Hout and M. Ernestus (2004), 'Pitfalls in corpus research', *Computers and the Humanities*, 38: 343-362.

Rude, S.S., E.M. Gortner and J.W. Pennebaker (2004), 'Language use of depressed and depression-vulnerable college students', *Cognition and Emotion*, 18: 1121-1133.

Scott, M (2004), *WordSmith Tools 4*. Oxford: Oxford University Press.

Scott, M. and C. Tribble (2006), *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Seale, C., S. Ziebland and J. Charteris-Black (2006), 'Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups', *Social Science and Medicine*, 62: 2577-2590.

Siegel, S. and N.J. Castellan (1988), *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.). New York: McGraw-Hill.

Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Spence, D.P. (1980), 'Lawfulness in lexical choice: a natural experiment', *Journal of the American Psychoanalytic Association*, 28: 115-132.

Spence, D.P., H.S. Scarborough and E.H. Ginsberg (1978), 'Lexical correlates of cervical cancer', *Social Science and Medicine*, 12: 141-145.

Swales, J. (1990), *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Tribble, C. (1998), *Writing Difficult Texts*. Unpublished PhD thesis, Lancaster University. Available from http://www.ctribble.co.uk/text/phd.htm.

Tribble, C. (2000), 'Genres, keywords, teaching: towards a pedagogic account of the language of project proposals', in: L. Burnard and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora.* Hamburg: Peter Lang. 75-90.

Tribble, C. (2001), 'Small corpora and teaching writing: towards a corpus-informed pedagogy of writing', in: M. Ghadessy, A. Henry and R. Roseberry (eds.) *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam and Philadelphia: John Benjamins. 381-408.

Zhang, H., C. Huang and S. Yu (2004), 'Distributional consistency as a general method for defining a core lexicon', in: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 26-28 May 2004. Available online from: http://data.cstr.ed.ac.uk/internal/library/proceedings/2004/lrec2004/.

On the phraseology of Chinese learner spoken English: Evidence of lexical chunks from COLSEC[1]

Naixing Wei

Shanghai Jiao Tong University

Abstract

Many previous theory-driven studies of phraseology basically focus on fixed and semi-fixed expressions which are structurally well-formed and semantically idiomatic. Altenberg (1998) is a landmark study of phraseology in a corpus-driven research paradigm, which tackles a much wider range of lexical sequences displaying complex structural and functional characteristics. The present paper, adopting Altenberg (1998) as a major referential framework, sets out to describe and discuss the structures and functions of lexical chunks in the Chinese Learners Spoken English Corpus*, with a view to characterizing the phraseological features of Chinese learner spoken English. The learner chunks are treated within three major structural categories, full clauses, clause constituents and incomplete phrases, within which pragmatic functions are categorized and examined. The paper concentrates on discrepancies and differences between learner and native speaker phraseologies. It has been found that, although the overall data distributions across major structural types show a similar tendency in the two sets of data, learners tend to use different chunks from native speakers for given meanings and functions. The study also reveals that learners more frequently use more types of recurrent sequences largely associated with the making of propositions, whilst using far fewer chunks (and using them much less frequently) which are basically associated with pragmatic functions. On the other hand, the learner interlanguage has developed its own characteristic ways of realizing pragmatic meanings, with specific unique evaluative chunks occurring frequently in a given situation. The paper also addresses the implications for second language acquisition research and second language pedagogy.*

1. Introduction

Phraseology has been a central focus of research in linguistics for a long time. Studies of this kind can be thought of as falling into two major paradigms: theory-driven studies and corpus-driven studies. Representatives of the former paradigm include, among others, Bolinger (1976), Pawley and Syder (1983), Nattinger (1988), and Cowie (1988), which, starting with their respective theoretical models, select a limited range of lexical sequences for study, referred to variously as linguistic prefabs, chunks, lexicalized sentences stems, lexical

phrases, and composites and formulae. (For a detailed review of these studies, refer to Wei 2002: 27-33). Different terms for more or less the same linguistic phenomena often reflect the researchers' differing theoretical perspectives and methodological considerations. Nevertheless, the theory-driven studies have several features in common: their point of departure is a theoretical model, within which a small number of lexical sequences is involved and discussed in detail; their principal criterion for identifying a sequence as phraseological is so-called 'psychological salience', whilst the frequencies of occurrence of sequences have not been seriously taken into account; and they largely focus on structurally well-formed entities (e.g. *I'm sorry to keep you waiting*) and semantically idiomatic phrases (e.g. *kick the bucket*). The corpus-driven research paradigm, by contrast, starts with data. The major criterion for identifying phraseology is probabilistic information or the frequency of occurrence of a sequence, rather than subjective psychological salience. Usually a very large number of sequences is examined and studies often deal with wider varieties of lexical sequence, for example, collocations, fixed and semi-fixed expressions, idioms, and incomplete lexico-grammatical sequences.

The corpus-driven model has brought about many a landmark study of phraseology with distinctive features. Sinclair and his colleagues at COBUILD initiated this type of research (see Sinclair 1987a).[2] And Sinclair (1987b: 320 and 1991: 110) formulated the Idiom Principle, which is of far-reaching significance for descriptive linguistics. Renouf and Sinclair (1991) shed light on the roles of discontinuous sequences in forming collocations. Biber (2004) noted that most lexical bundles are structurally incomplete and semantically non-idiomatic. Altenberg (1998) is a milestone in the corpus-driven study of phraseology, which underscored, among other things, the ubiquity of Recurrent Word Combinations, their formal flexibility, semantic transparency and pragmatic conventionality. Studies in learner corpora addressed the phraseological features of foreign learners' English writings (De Cock et al. 1998; De Cock 1998; Pu 2003). Wei (2004) is a preliminary report on the various language features and discourse patterns as shown in the *Chinese Learners Spoken English Corpus* (COLSEC), which examined part of the lexical chunks recurring in the corpus. So far, however, systematic studies of the phraseology of English as a Foreign Language (EFL) learners' spoken English, based on relatively large quantities of data, are few and far between. What are the major characteristics of foreign learner phraseology in English conversation, in terms of their formal and functional features? To what extent are these features similar to, or different from, those of native speakers? What, if any, unique patterns and features has learner interlanguage developed for realizing meanings and functions in spoken communication? These questions merit a thorough and methodical investigation in corpus linguistics, which may provide new insights for understanding phraseology, and, in particular, for understanding linguistic choices and meaning/function realizations in second language communication.

In this paper, I set out to investigate and describe the phraseological features of Chinese learner spoken English, on the basis of evidence from

COLSEC. The purpose of this article is twofold. First, I wish to present an overview of the gamut of typical lexical chunks used by Chinese learners in English conversation through a comprehensive documentation of data, discussing their formal and functional characteristics. Second, I wish to uncover part of the important formal and functional features of the learner phraseology by comparing chunks used by native speakers and learners. I will adopt Altenberg (1998) as my referential framework for the investigation. In the sections that follow, I shall first give a brief introduction to COLSEC and describe my research methods. Secondly, I shall present the overall data distribution of the Chinese learner lexical chunks of various lengths and in different structural and functional categories. I then devote several sections to a discussion of the formal and functional characteristics of chunks in the major categories. Similarities and differences between learner chunks and native speaker chunks will be spelt out in both quantitative and qualitative terms. And I shall argue that pragmatic, or evaluative, chunks are a serious area of weakness in learner speech. Lastly, I summarize the major findings of the present study and address their implications.

2. COLSEC and research methods

The present study uses data from COLSEC for investigating the phraseological features of Chinese learner spoken English. The COLSEC project was funded by the Chinese National Social Science Research Foundation, which aims to investigate characteristics of Chinese learners' spoken English and provide feedback for college English learning and teaching in China. The project was launched at Shanghai Jiao Tong University in the year 2000 and completed in 2005, with a resultant 700,000-word corpus of learner spoken English. Raw materials for COLSEC are the test 'episodes' from CET-SET (Spoken English Test component of the College English Test),[3] which is administered nation-wide twice a year. Each episode in the test consists of three sections, namely, an interview section, in which the examiner (a teacher) and the examinee (a student) perform question-answer tasks concerning the examinee's academic study, campus life and other familiar topics; a discussion section in which three examinees discuss and debate certain social issues of common interest; and, finally, a further discussion section in which the examiner and an examinee re-discuss particular questions of special interest which have just been discussed in the previous section. All the examinees' performances are video-recorded and graded according to a set of criteria (Yang 1999). The episodes are sampled according to the examinees' grades in the test, their topics in discussion and their academic specialities in the university. The selected video-recorded test episodes are then transcribed and annotated in terms of discourse features, including conversational turns, fillers, interruptions, repetitions and pauses; and phonological features, including various types of mispronunciation, misplaced stresses, non-verbal sounds and indistinguishable sounds.[4] Table 1 presents the overall statistics of the corpus.

Table 1: Overall statistics of COLSEC

	Numbers		Numbers
Test episodes	302	4-letter words	142143
Conversation topics	39	5-letter words	76166
Tokens	723299	6-letter words	44001
Types	9192	7-letter words	38718
Type/token ratio	1.27	8-letter words	23228
Std. type/token ratio	28.44	9-letter words	18368
Av. word length	4.01	10-letter words	11643
Sent. length	12.79	11-letter words	5342
Std. sent. length	12.13	12-letter words	2672
1-letter words	38178	13-letter words	1370
2-letter words	163778	14(+)-letter words	360
3-letter words	157267		

The method employed in the present study is as follows. I temporarily define a lexical chunk as 'a continuous lexical sequence of different lengths which has occurred frequently in COLSEC'. The term 'lexical chunk', in essence, refers to the same linguistic phenomenon as a 'recurrent word combination' defined in Altenberg (1998). But, as I see it, the term 'recurrent word combination' sounds very much like collocation, which actually differs in many ways. Thus, I prefer to use the term 'lexical chunk' in the present investigation. Occasionally the term 'lexical sequence' is used interchangeably with 'lexical chunk'. As a first step, I used the wordlist function of WordSmith Tools to several times compute and obtain word frequency information in COLSEC. Each time I varied the 'cluster length setup' of WordSmith Tools, so that 2-word, 3-word, 4-word, 5-word and 6-word wordlists could be obtained successively. In so doing, I essentially cut up the whole corpus into sequences of different physical lengths. As a result, lexical sequences of 2-word lengths through to 6-word lengths were obtained. Secondly, I delimited the lexical chunks for study by setting cut-off frequencies for sequences of different lengths to remove the less frequent sequences from the list: the cut-off frequencies for 2-word through to 6-word sequences were 15, 15, 10, 10 and 10, respectively. These frequency criteria were arbitrarily used for the purpose of extracting a batch of frequent chunks from a formidably large quantity of data for me to focus on. As a third step, I manually removed some frequent sequences out of the lists, as these consist of non-verbal sounds and repetitions which are very often signs of stuttering and dysfluency, for example, *mm in my*, and *mm you can can*. Fourthly, in the qualitative analyses, I adopted Altenberg's (1998) framework to describe and discuss the formal and functional features of the learner chunks. As with Altenberg (1998), this study concentrated on the features of chunks of length of at least three words, since a large number of 2-word chunks are compound nouns (e.g. *college education*, *computer science*), and those denoting grammatical information (e.g. *he will*, *we are*). A major focus of

the present study is on the unique features of learner chunks, for which in-depth analyses will be carried out with reference to related previous studies.

3. Overall data and structural types of chunks

Adopting the above method as defined, the present study obtained 2,702 different lexical chunks of various lengths, which are referred to as 'chunk-types'. The overall frequency of occurrence of these chunks in COLSEC is 92,635, which are called 'chunk-tokens'. Table 2 shows the overall distribution of chunks. Figure 1 is a graphic representation of the frequencies of chunks of different lengths in the corpus.

Table 2: Overall distribution of chunks in COLSEC

Chunks	Types	Tokens
3-word chunk	1382	63766
4-word chunk	866	20815
5-word chunk	258	5084
6-word chunk	196	2970
Total	2702	92635

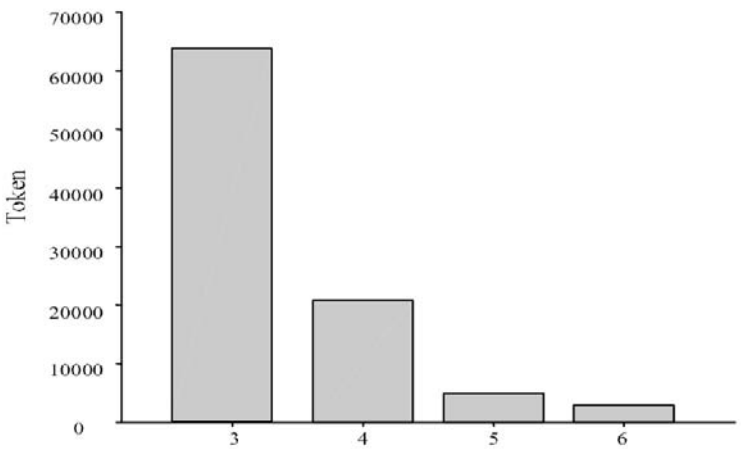

Figure 1: Frequency of chunks of different lengths in COLSEC

As shown in table 2, 3-word chunks make up the largest proportion of all chunks, whether in terms of types or in terms of tokens. 4-word chunks are the next most frequently used, whilst the numbers of types and tokens of 5-word chunks and 6-word chunks are sharply reduced. The distribution of data with respect to the

chunks of four different lengths shows a similar tendency to that of Altenberg (1998). However, it must be pointed out that these represent only a small proportion of the lexical sequences occurring in COLSEC, as I have adopted a very conservative approach in order to reduce the data to a manageable size.

Adopting the framework of Altenberg (1998), I divided the chunks in COLSEC into three structural types: full clauses, clause constituents and incomplete phrases. Table 3 shows the frequencies of these major structural types as well as their subtypes.

Table 3: Frequency distribution across structural types

Structural class	Frequency	Percentage
Full clauses	4259	4.60
Independent clauses	3861	4.17
Dependent clauses	398	0.43
Clause constituents	80760	87.18
Multiple clause constituents	63545	68.60
Single clause constituents	17215	18.58
Incomplete phrases	7616	8.22
Total	92635	100.00

As shown in table 3, clause constituents take the lion's share of the overall frequency of chunks, with their occurrences accounting for 87.18 percent of the total; incomplete phrases are the second most frequently occurring sequences, but these frequencies are substantially lower. By contrast, full clauses have the lowest frequency of occurrence, only accounting for 4.6 percent of the total instances.

This data profile is, to a great extent, similar to the data distribution across the structural types reported in Altenberg (1998: 103).[5] These appear to suggest that clause constituents are the single most important building blocks for ongoing discourse, whether in native speaker or in learner spoken communication. But since clause constituents do not represent a complete structural or semantic unit, they have been ignored in theory-driven studies of phraseology. However, with their high frequency of occurrence, these chunks merit a detailed investigation in corpus linguistics.

At the same time, closer observation of the two sets of data shows that there are discrepancies between them. Clause constituents in COLSEC make up a higher proportion of the total than their counterparts in Altenberg (1998: 103) (87.18 percent versus 76 percent). As a consequence, the other two structural types in COLSEC accounted for lower proportions than in the case of Altenberg (1998). This will be returned to in more detail in 4.1.

4. Full clauses and their pragmatic functions

Full clauses have a complete subject-predicate structure, which comprise two subtypes: independent clauses and dependent clauses. In phraseological studies,

full clauses are lexical sequences which are at the highest syntactic level. From the functional perspective, these are highly conventionalized formulae associated with specific contexts of situation. Due to the demands of real-time conversation, speakers often choose relevant ready-made full clauses to fulfil various functions in a given context of situation. In what follows, I describe and discuss the major pragmatic functions of independent clauses and dependent clauses in COLSEC.

4.1 Major functions and formal features of independent clauses

4.1.1 Major functions of independent clauses

A close observation of the data reveals that there are both similarities and differences between the independent clauses used by learners and native speakers. In functional terms, these independent clauses can be divided into twelve categories, namely, phatic expressions, thanks, questions, acknowledgement, agreement, disagreement, positive polar, negative polar, disclaimer, self-introduction, statement and others. Among them, six categories are unique in COLSEC, including phatic expressions, questions, disagreement, statement, self-introduction and others. The category defined as 'reassuring' in Altenberg (1998: 104) is absent in COLSEC. For reasons of space, table 4 below presents just a sample of the most frequently occurring chunks for each functional category.

Phraseologically, chunks in table 4 are basically semi-fixed formulae, allowing lexico-grammatical variations, or permitting expansions and reductions, to varying degrees. For instance, *nice to meet you* can be varied to *nice to have met you*; *thank you very much* can be reduced to *thank you*, and can also be expanded to *thank you very much indeed*, or varied to *thank you ever so much*, etc; and *I beg your pardon* can be expanded to *I humbly/do beg your pardon*, etc. In other words, these clauses are lexicalized to different extents. All the chunks are semantically transparent except *I beg your pardon*. It seems to be the case that learners choose them to realize specific speech act functions as required by contexts of situation. For instance, phatic expressions are used to greet examiners and fellow examinees at the start of the test to show politeness; acknowledgement is used to confirm the receipt of information; agreement is used for expressing approval of other's viewpoints; whereas disagreement is used for rejecting a differing view. In actual fact, the six unique functional categories of chunk occur in COLSEC because they are required in the fulfilment of specific communicative tasks in the CET-SET test. Nevertheless, several discrepancies between the native data and the learner data can be noted.

First, in quantitative terms, learners used far fewer instances of full clauses than native speakers. Full clauses only accounted for 4.6 percent of the occurrences of all chunks, while they accounted for ten percent in the case of Altenberg (1998), i.e. over twice that seen in the learner data. This appears to suggest that the learners are not particularly adept at manipulating these ready-made chunks for pragmatic functions, for the full clauses are chiefly the conventionalized chunks which are closely associated with specific pragmatic functions in a given

situation. The automatic selection and appropriate use of these pragmatically specialized formulae (Cowie 1988) saves time for encoding and is conducive to fluency of speech. The learners' weakness in manipulating these chunks may partly account for their non-nativeness and dysfluency in speaking.

Table 4: Major functions of independent clauses in COLSEC

Functions	Instances	Freq.
Phatic expressions	nice to meet you	60
	glad to meet you	30
Thanks	thank you everybody	85
	thank you very much	55
Questions	what's your opinion	113
	do you think so	36
	what about you	31
Acknowledgement	ok (oh, now, yes) I see	15
Agreement	I think so	181
	I agree with you	175
	quite agree with you	26
Disagreement	I don't think so	112
	I don't agree with you	21
Positive (polar)	oh yes (yes)	48
	yes of course	40
Negative (polar)	no no (no)	17
Disclaimer	I don't know	73
Self-introduction	I am a junior	77
	let me introduce myself	21
Statement	it's very important	75
	I like it	54
Others	I am sorry	53
	I beg your pardon	26
	everything has two sides	20

Second, learner chunks betray unnaturalness as compared with the chunks of native speakers. Although overall quantitative distributions across the three major structural types are similar in the two sets of data, many learner chunk-types are very different from those of native speakers. For one thing, full clause chunks in table 4 are much longer than those reported in Altenberg (1998: 104). In the latter case, frequently occurring sequences are *that's right*, *yes it is*, *yes I have*, *no I don't*, *oh no no*, etc. (Altenberg 1998: 104). In contrast, learners very often use longer clauses, as shown in table 4. These longer clauses may sound jarring and unnatural to the native ear. Taking the functional category 'agreement' as an example: the typical native speaker chunk is *(yes) that's right* (Altenberg 1998: 104), but the most frequent learner chunks include *I agree with you* and *quite agree with you*. Unnaturalness will be returned to in more detail in 4.1.2.

Thirdly, intercultural influences can be noticed in learner chunks. An example of such influence can be identified in the formulaic sequence *everything has two sides*. Semantically, this formulaic sequence bears some resemblance to the English cliché *two sides of the same coin.* But learners, more often than not, use the former sequence rather than the latter because of the influence from Chinese culture: according to a prevalent view in Chinese philosophy, the whole universe and the human world are composed of two types of element, *yin* and *yang* (negative and positive elements).[6] This philosophical view heavily influences the language use and has resulted in a Chinese saying, which can be literally translated into *everything has two sides* in English, which, then, has become a frequent chunk in the learner English speech, as shown in the following extract from COLSEC.

(1) <sp1> Em, in my opinion, I think a family with one child is all right, is better than with many children. Er, for ... we were, we are the only one child in the family. Our parents give us the best care. </sp1>
<sp2> I agree with you. Em, we can have very good education. </sp2>
<sp3> Well, ... everything has two sides, you know. Just think, will you feel lonely when you are alone at home and there is nobody accompany with you? You've no brother or sister to play with. </sp3>

In the above extract, three students are discussing the one-child-family programme currently adopted by the Chinese government as a fundamental state policy for family planning. "sp1" and "sp2" talk about the positive side of the one-child-family, while "sp3", by using the chunk "everything has two sides", turns to its negative side.

4.1.2 Forms and functions of independent clauses

4.1.2.1 Use of core forms and their related variant expressions

Altenberg (1998) distinguishes between two forms of independent clause: the core form and the expanded variants. According to him, the latter is more frequently used than the former in real communication. For instance, in the *London-Lund Corpus*, the core form *I see* for 'acknowledgement' occurs 58 times, but its related expanded variants, *oh I see*, and so on, are twice as common as the core. In other words, the core is the marked form, while expanded variants are unmarked forms, which are more natural and appropriate in normal contexts of situation (Altenberg 1998: 107-109). This is an interesting feature of independent clauses used in spontaneous speech. How, then, do Chinese learners use these two forms of independent clause? Do they more frequently use the core, or the expanded variants? The present study has investigated the core forms and the expanded forms of, among others, the acknowledgement chunk *I see* and the positive polar *yes*.

There are 15 instances of *I see* in COLSEC (see concordances in appendix). Among them, the core form *I see* occurs five times, *now I see* once, *oh*

I see four times, *ok I see* twice, and *yes (yeah) I see* three times. All in all, one-third of the occurrences are in the marked core form while two-thirds are in the unmarked expanded forms. Such a distribution is similar to that of native speakers, in the case of *I see*, which roughly shows that learners have acquired the use of different forms of this acknowledgement chunk.

But the uses of the positive polar *yes* and other chunks show a different picture. The data distribution of the positive polar *yes* is as follows: *oh yes*: 48 times; *yes but*: 36 times; *yes ok*: six times; *yes right*: three times; *yes*: 1,077 times. To put it simply, the core form occurs 1,077 times, which is eleven times as frequent as the expanded variants whose total occurrences add up to 93.

Using Altenberg's line of thinking, it could be said that, while the marked and unmarked forms of some independent clauses have been used in reasonably native speaker-like proportions in COLSEC, many clauses in the learner data differ a great deal from those in the native data, with marked and unmarked forms occurring in remarkably different proportions.

4.1.2.2 Different forms for different functions

However, I cannot totally agree with Altenberg (1998) in regarding the related expanded forms as the variants of a so-called core form. My position is: a related long chunk can be thought of as the expanded variant of a short chunk only when the two forms perform one and the same pragmatic function in a given context of situation. For example, the long chunk *thank you very much* can be regarded as the expanded variant of the short chunk *thank you* (Altenberg 1998: 106), because both perform one and the same function, namely, that of expressing thanks. In terms of discourse analysis, whether it is a related long sequence, such as *thank you very much*, or the short sequence, such as *thank you*, that is being used, it constitutes one discourse move or act in the given context (refer to Tsui 1994). But not all the instances discussed in Altenberg (1998: 107-108), including the related expanded forms of *I see*, *now I see*, *oh I see*, *ok I see*, etc. can be reasonably taken as the expanded variants of *I see* in the strict sense of the term as I have just expounded. The first elements of these expanded forms, *now*, *oh*, and *ok* are, in essence, different discourse particles in their own right, which perform a variety of discourse functions in given contexts (Aijmer 2002). In other words, in such cases, a related long chunk can be a combination of two different utterances, rather than being the expanded variant of the so-called core form.[7] Let me examine the sequence *yes but* in a stretch of conversation from COLSEC presented in (2).

(2) <sp1> Also I'd like to say that there are some graduates from colleges. They are masters, even doctors. … mn, but, er …, but their ability to deal with things, I think, er, is even lower [Wl-n] than those [Wth-n] people who didn't enter the university. </sp1>
<sp3> Yes, but there is another fact, er ... another thing that I'm worrying about, that is, everything is changing, … </sp3>

The above stretch of talk reveals that two students are talking about the practical ability of college graduates. "sp1" says that the ability of some graduates is actually lower than those who have not received higher education. After "sp1" finishes his turn of speaking, "sp3" uses the sequence *yes but* to start his turn. But a close examination of the context shows that the positive polar *yes* is used to express the current speaker's approval of "sp1"'s view. However, the connector *but* is used by the current speaker to launch a new discourse move in which he discusses the topic from a different angle. In discoursal terms, *yes* and *but* perform two different functions and belong to two different discourse moves. Out of these considerations, it is better to think of the sequence *yes but* as the combination of two different utterances or moves. In other words, *yes but* can be a chunk different from the single *yes* in both form and function. Therefore, when learners used expanded forms and non-expanded forms concerning a chunk differently from native speakers, it may suggest that different discourse functions have been realized in learner talk.

The above discussion seems to boil down to a few points as regards the formal and functional features of independent clauses in COLSEC. First, some independent clauses in learner conversation seem to bear formal features similar to those of their counterparts in native speaker data, considering that similar proportions of the marked forms and unmarked forms of some chunks occurred in the two sets of data. But this fact can also mean that the functions realized in the learner corpus are more or less the same as those in the native speaker data. Second, many independent clauses in learner conversation seem to bear formal features remarkably different from those of their counterparts in native speaker data, as far as the proportions of the marked forms and unmarked forms of the chunks are concerned. But where learners have used expanded forms and non-expanded forms concerning a chunk differently from native speakers, it may suggest that both the forms and the functions of learner discourse differ from those of native speakers, and this issue needs to be investigated further in future studies. Third, in general terms, marked and unmarked forms relate to pragmatic concerns of the appropriateness of forms in context. This points to the necessity of giving more weight to the pragmatic appropriateness of forms in EFL learning and second language acquisition (SLA) research.

4.2 The major functional and formal features of dependent clauses

Dependent clauses are divided into three functional types: comment clauses, indirect conditions and apposition markers. Comment clauses are used to quote shared background information and facts, or to introduce speakers' viewpoints. Indirect conditions are politeness formulae, mainly including *if I may*, *if you like*. Apposition markers reformulate part of the utterance. Table 5 below shows the data of dependent clauses in COLSEC with frequencies above the cut-off point.

Data in table 5 reveal both similarities to, and differences from, the native speaker data. The comment clause *as you know* and the apposition marker *that is to say* appeared frequently in both sets of data (Altenberg 1998: 109). But all the comment clauses in table 5 are introduced by the conjunction *as*, followed by a

personal pronoun (*you* or *we*) plus a verb (most often the verb *know*), except for the semi-fixed phrase *as far as I'm concerned*. Other frequent comment clauses in the native data (Altenberg 1998: 109), for instance *I should like*, *as it were*, failed to occur at all in the learner data. Formally, these *as*-introduced clauses are largely semi-fixed expressions, permitting variations. Semantically, they are transparent and close to one another in meaning. Pedagogically, they are phrases learned early in Chinese learners' English courses, and thus familiar to them. In addition, it should also be noted that there is no occurrence of indirect conditions expressing politeness in table 5, which might mar the pragmatic quality of discourse. It seems to be the case that learners are capable of using some comment clauses, but their choice of such sequences is limited in variety and range.

Table 5: Functions and data of dependent clauses

Functional types	Instances	Freq.	Functional types	Instances	Freq.
Comment clause	as we know	73	Comment clause	as I know	28
	as we all know	68		as we can see	16
	as you know	51		as far as I know	16
	as far as I'm concerned	42		as you see	15
	(just) as I said	33	Apposition marker	that is to say	35

5. Clause constituents and their pragmatic functions

Clause constituents do not stand-alone but act as components of a clause. There are two subtypes of clause constituents: multiple clause constituents and single clause constituents.

5.1 Multiple clause constituents

Multiple clause constituents cover a wide range of lexical sequences. They make up 68.60 percent of all the chunk-tokens, which is the highest ratio for all the chunk-subtypes. Following Altenberg (1998), I have divided multiple clause constituents into two major types: 'thematic springboard' and 'propositional core', according to their linear position in the clause. The former type denotes given information and is further divided into three interconnecting categories of sequence: frames, onsets and stems, while the latter type conveys new information and covers rhemes, tails and transitions.

Because frames and stems serve as the most important components in the thematic springboard, while onsets commonly overlap with these two categories, and also because rhemes are the central component of the propositional core, while tails and transitions act as additional components, I shall, in what follows,

first concentrate on discussing the features of frames, stems and rhemes in COLSEC, and then briefly generalize the functions of multiple clause constituents.

5.1.1 Frames in COLSEC

Frames are in pre-subject position and generally consist of combinations of a sentence connector and a discourse particle or a discourse signal. Table 6 shows the twenty most frequently occurring frames in COLSEC.

Frames are a useful means of starting an utterance and indicate the directions in which discourse information will unfold. For this reason, frames are referred to as "utterance launchers" by Biber et al. (1999: 1073-1078) and as "clause openers" by Altenberg (1998: 113). As shown in table 6, whether in terms of type or token, the most frequently occurring connectors in frames turn out to be *and*, *so*, *but*. Then the connector *because* is relatively frequently used. The connectors *then*, *also* and the discourse particle *well* are the least frequently used, each of which has occurred in one chunk only, respectively *then I think*, *also I think* and *well I think*. For the second component of the frame, the commonest lexical means are the discourse particles *I think*, *you know*, the discourse modal *of course* and discourse signals *in my opinion*, *first of all*, *in fact* and *on the other hand*.

Table 6: Most frequently occurring frames in COLSEC

Frames	Freq.	Frames	Freq.
so I think	1026	then I think	25
and I think	965	and so I think	25
but I think	556	and of course	21
because I think	230	and also I think	19
because you know	75	but in my opinion	18
and you know	60	and first of all	16
also I think	39	but in fact	15
and I think that	32	so in my opinion	14
but you know	30	and in my opinion	11
well I think	26	but on the other hand	11

There seems to be a co-selection between the first and the second frame components. The causal connector *so* and the discourse particle *I think*, as well as the discourse signal *in my opinion*, are mutually selected to generalize opinions and proposals; the adversative connector *but* and *I think*, as well as *on the other hand* and *in my opinion*, are mutually selected to present disagreement and different facts; the connector *and* and various lexical means are combined to present a variety of discourse information. The discourse particle *you know* often serves the role of narrowing the affective distance between speakers or reinforc-

ing a certain fact; thus, *you know* is combined with *and*, *but* to highlight information.

It must be pointed out, however, that there are differences between the frames in learner English and native speaker English. For one thing, the first components of frames in table 6 are almost exclusively sentence connectors of various kinds, with the exception of the frame *well I think* in which the first component is a discourse particle. In contrast, in native speaker data, the first frame components include both sentence connectors and a variety of discourse particles, such as *well*, *I mean*, etc. Particularly, the discourse particle *well* is often selected by native speakers to start a frame sequence. As a result, frame sequences, such as *well I mean*, *well you see*, *well of course* and *well you know*, are reported as commonly occurring in Altenberg (1998: 112); but they failed to occur at all in COLSEC. However, their failure to occur in the learner corpus does not mean that the independent item *well* has not occurred frequently in it. My data reveal that *well*, as a discourse particle, occurred 236 times in COLSEC, a standardized frequency of 3.4 per 10,000 words. But, characteristically it occurs on its own, not co-occurring with the other pragmatic devices *you know*, *you see*, *I mean* and *of course* to start a conversation turn. The data also show that these other pragmatic devices occurred in the learner corpus with different frequencies as follows: *of course*: 151 times; *you know*: 87 times; *you see*: 26 times; *I mean*: 22 times. But, in similar ways, they characteristically do not come together to start an utterance, or launch a conversation turn. As a consequence, the above-mentioned *well*-introduced frames and a variety of others (cf. Aijmer 2002) do not appear in learner English. Taken together, these features may suggest that there exists a set of differences between learner English and native speaker English, as regards the forms and functions of linguistic means, with which to launch utterances and conversation turns, a topic which merits more thorough investigation in future studies.

5.1.2 Stems in COLSEC

Stems contain a subject and a verb, and lack any following objects and complements. Altenberg (1998) divided stems into five functional types: epistemic, existential, reporting, interrogative and others. Table 7 below shows a sample of the most common stems in COLSEC, which provides a rough picture of the structures and functions of the major stems used by learners.

These stems carry the most important thematic elements and the given information in the discourse. A noticeable difference between the learner data and the native speaker data lies in the use of epistemic stems. In COLSEC, the most frequently occurring epistemic stems are the *I + epistemic verb* sequences. In addition to *I think*, *I believe* and *I guess*, there are also *I suppose*, *I assume*, *I thought*, which are not listed in table 7 because of their low frequencies. The frequent epistemic stems, *it seems that*, *it seems to me that*, *I would have thought that*, reported in Altenberg (1998: 114), did not appear in COLSEC. But learner data for existential, reporting and interrogative stems are more or less the same as for native speakers. However, there is a particular stem sequence, *it is very*, in the

subcategory of 'others', which is worth discussing. The sequence *it is very* occurred 291 times, and its variant *it's very* occurred 479 times. The sequence is often followed by an adjective: *good*, *important*, *convenient*, *bad*, *interesting*, *helpful*, *useful* and *necessary* are the most frequent items. A close examination of the evidence reveals that the *it is very ADJ* sequence is attitudinal in nature (see concordances in appendix). The sequence often co-occurs with the discourse particle, or the epistemic stem, *I think*, to present comments and judgments. The fact that the sequence, particularly its combination with *I think*, occurred frequently in COLSEC suggests that it is an important means for learners to express evaluative meanings. That is to say, learners, in most cases, tend to use this sequence to comment on and express a general view of things. This feature for expressing attitudinal meanings is unique to learner speech.

Table 7: Most frequently occurring stems in COLSEC

Type	Instance	Freq.
Epistemic	I think (that) [8]	1507
	I don't think (that)	280
	I believe (that)	82
	I guess (that)	43
Existential	there are (many, some)	549
	there is a/n	187
Reporting	he/I/she/they said (that)	162
	I have said	28
Interrogative	do you think	1021
	what do you think	238
	do you like	93
	have you ever	93
	do you have	71
	can you say	62
	do you know	43
Others	my name is	1477
	you will have	622
	I'd like to	581
	I want to	429
	it is very[9]	291

5.1.3 Rhemes in COLSEC

Rhemes consist of verbs and their complements, which partly overlap with stems. Rhemes act as the most important element of the propositional core. Unlike thematic springboard chunks, rhemes contain at least one open-class word, displaying high degrees of lexical variation. Table 8 below shows the 20 most frequently occurring rhemes in COLSEC.

Table 8: Most frequently occurring rhemes

Rheme	Freq.	Rheme	Freq.
have a lot	164	help each other	67
take care of	123	get along with	62
take part in	116	save money for	56
is the best	110	try my best	56
is the most important	108	is good for	50
is very good	94	pay more attention	49
is more important	89	spend a lot	48
find a job	85	help us to	47
do a lot	80	learn a lot	47
make friends with	69	communicate with others	45

Altenberg (1998) found that thematic springboard sequences are mostly highly recurrent, whereas sequences in the propositional core are rarely recurrent. As a result, he did not report any actual rheme sequences. As observed in table 8, however, a good many rhemes in COLSEC are highly recurrent, which is a prominent feature of learner English speech. Some chunks in table 8 obviously reveal topic constraints, for instance *find a job*, *make friends with*, *save money for*, *spend a lot.*[10] A few others are not closely related to topics, but act as a common means of expressing personal opinions and attitudes, for instance *is the best*, *is very good* and *is more important.* Many rhemes are closely associated with the making of propositions. That they are highly recurrent can be attributed to the limitedness of learner English proficiency. Data in table 8 reveal that many chunks are fixed and semi-fixed expressions which learners acquire at very early stages of their English learning, for example *take care of*, *take part in*, *help each other*, *get along with*, *try my best*, *pay more attention*, and *help us to*.

5.2 Single clause constituents

Single clause constituents are basically complete phrases, involving both fixed and semi-fixed phrases as well as free combinations of high flexibility. Data show that single clause constituents in COLSEC not only cover the functional categories of vagueness tags, intensifiers/quantifiers, discourse connectors, temporal expressions, and spatial expressions, as identified in Altenberg (1998), but also include two unique categories: purposive expressions and nominal expressions, as presented in table 9 below.

Compared with multiple clause constituents, single clause constituents are phraseologically more interesting. As pointed out above, these sequences are lexicalized to a greater extent. In terms of grammatical structure, most single clause constituents are prepositional phrases functioning as adverbials. Functionally, vagueness tags, intensifiers/quantifiers, and discourse connectors help to realize interpersonal and textual meanings in the Hallidayan sense (Halliday 1985), and are pragmatic in nature. Whilst temporal expressions, spatial

expressions, nominal expressions and purposive expressions are largely proposition-making-oriented.

Table 9: Highly recurrent single clause constituents

Type of element	Instances	Freq.
Vagueness tags	(and) and so on	123
	something like that	16
	or something like that	14
Intensifiers/quantifiers	more and more	447
	to some extent	40
Discourse connectors	first of all	406
	in my opinion	279
	as for me	76
	at the same time	54
	on the other hand	48
Temporal expressions	in my spare time	139
	in the future	250
	in the morning	62
	for a long time	20
Spatial expressions	in the society	117
	in the university	107
	on the Internet	99
	in the world	75
Nominal expressions	fake and inferior products	96
	benefits and problems	93
	science and technology	38
	the outside world	21
Purposive expressions	for the future	74
	to stay healthy	57
	for future use	54
	to make friends	54

Data in table 9 reveal remarkable differences between the single clause constituents in COLSEC and those reported in Altenberg (1998). First, vagueness tags have been seriously under-represented in COLSEC. In comparison with the native data, learners almost exclusively relied on the chunk *and so on* for expressing vague meaning, with the other two chunks, *something like that*, *or something like that*, much under-used, and another four chunks, *and all that*, *sort of thing*, *that sort of thing*, *and things like that*, never used at all.[11]

In recent years, vagueness tags have received considerable attention in pragmatic studies. In the literature, this research topic has been referred to variously as 'vague category identifiers' (Channell 1994), 'clause terminal tags' (Dines 1980), 'vagueness tags' (De Cock 2000), and 'utterance-final tags' (Aijmer 2002). Vagueness tags are commonly attached to the end of clauses or

phrases to denote that the information represented in the preceding clause or phrase is merely one of the instances of a vague or imprecise category. De Cock (2000) is an in-depth comparative study of the use of vagueness tags by native speakers and by learners whose mother tongue is French. She examined uses of more vagueness tags, but came to a similar conclusion that learners significantly under-used most vagueness tags compared with native speakers. Therefore, it seems to be true that vagueness tags are weak points for all EFL learners, irrespective of their native language backgrounds.

The major function of vagueness tags is pragmatic in nature. Due to the pressure in real-time speech and limited knowledge of the world, it is impossible to convey precise information at times. More importantly, for various purposes, such as being polite and cooperative, speakers may also deliberately convey information imprecisely. This has led to the extensive use of vagueness tags in native speaker speech. Although these linguistic devices sound empty in semantic terms, they are a natural and strategic part of speech. The significant under-use, and even zero-use, of vagueness tags may cause learner speech to sound more like written communication, or lacking in politeness and cooperativeness.

Second, the whole category of 'qualifying expressions' in Altenberg (1998: 117), including *more or less*, *in a way*, *in a sense* and *on the whole*, is absent in COLSEC. Intensifiers/quantifiers are much less used: the recurrent chunks reported in Altenberg (1998: 117), such as *the whole thing*, *a bit more*, *a lot more*, *a little more* and *the whole lot* are missing completely. Both qualifying expressions and intensifiers perform discoursal and pragmatic roles, for example reinforcing cohesion and attitude. The significant under-use and even zero-use of these categories of sequences may affect the pragmatic quality of the learner discourse.

Thirdly and finally, the frequent occurrence of purposive expressions and nominal expressions is a unique feature of the learner data. To a greater extent, these sequences tilt toward the making of propositions. With all these features taken together, it could be argued that the learner discourse is biased toward the making of propositions while manifesting weakness in its pragmatics.

6. Incomplete phrases

Along the lines of Altenberg (1998), I categorized the incomplete phrases in COLSEC into three types. The first is phrasal fragments, the items of which consist wholly of function words (e.g. *all of the*) or of function words plus one or two content words (e.g. *a chance to*). These fragments are not well-formed grammatically and are not psychologically salient either. They are, however, capable of generating complete phrases. Their combination with words of various kinds will constitute well-structured sequences, for instance *all of the time* and *one of the factors*. The high frequencies of these fragments in the corpus indicate that they are a useful means of constructing discourse and merit discussion in the context of phraseology.

Table 10: Phrasal fragments and semi-/fixed phrases

Phrasal fragments (1) (function words only)	one of the	81
	most of the	66
	such as the	56
	all of the	30
	some of the	24
Phrasal fragments (2) (with content words)	a good way to	68
	with the development of	66
	a chance to	22
	in the use of	12
Fixed and semi-fixed phrases	in order to	96
	to see if you	26
	according to the	23
	in front of	20
	at the end of	19

Table 11: Commonest collocational frameworks

Frameworks	Instances	Freq.
as + ADV + as	as far as	93
	as well as	21
	as much as	19
	as long as	15
a + N + of	a lot of	1158
	a kind of	66
	a number of	25
	a part of	16
	a waste of	15
	a form of	13
the + N + of	the name of	292
	the number of	249
	the topic of	244
	the development of	172
	the use of	129
	the importance of	35
	the problem of	27
others	this kind of	128
	all kinds of	79
	such kind of	35
	one way of	27
	various kinds of	25

The second type is fixed and semi-fixed phrases of various kinds, such as *in order to*, *to see if you*. The third type is the kind of sequences of the collocational frameworks defined by Renouf and Sinclair (1991). These are very common discontinuous sequences made up of function words on either side and enclosing characteristic content words, as in, for instance, *a* + *lot* + *of* and *the* + *importance* + *of*. Table 10 (above) shows the data for phrasal fragments and fixed and semi-fixed phrases, and table 11 (above) presents the data for collocational frameworks.

In general terms, the types of incomplete phrases used by Chinese learners are roughly similar to those of native speakers. On the other hand, there are noticeable discrepancies between the two sets of data. Many sequences, including *one of the*, *as far as*, *a lot of*, *a kind of*, *the importance of*, *the development of*, appear much more frequently in COLSEC than in the native data, whereas common sequences in the native data, such as *a sort of*, *a bit of*, *a couple of*, and *the whole of the*, are absent in COLSEC (cf. Altenberg 1998: 119-120). Much needs to be investigated concerning uses of these sequences by the learner; however, at this point, I will concentrate on the sequence *with the development of*.

While this chunk rarely occurs in any native speaker corpus, it occurred 66 times in COLSEC.[12] What is unique and remarkable about this chunk is that it has developed its characteristic formal and functional features, serving important discoursal and pragmatic roles in the learner speech. Formally, it almost always appears in clause-initial positions, or follows immediately on from the discourse particles *I think* and *you know*, etc. (Aijmer 2002) to start an utterance. Pragmatically, it shows an obvious positive semantic prosody (Louw 1993), co-occurring, on the right, with words of positive connotation, such as *country*, *economy*, *society*, *science*, *technology*, which refer to things that have been developing very fast in China. The chunk has become, to varying degrees, a device for the learners to air positive views and judgements. The positive views are often expressed or implied in the wordings on the right of the chunk, as shown in the following citations from COLSEC.

(3) I think, with the development of technology, our high-tech products will become more competitive.
(4) With the development of our society, people can enjoy more freedom to go abroad for education.
(5) I think, with the development of economy, people had their different attitudes towards how to spend money. Unlike the old generation, many of us young people just want to spend all the money in our pockets.

In citations (3) and (4), "our high-tech products will become more competitive" and "people can enjoy more freedom to go abroad for education" are overt expressions of the speaker's positive evaluation and judgement. But in citation (5), what is conveyed is more of a covert or implied positive evaluation on the "different attitudes towards how to spend money": the larger context reveals that three students are discussing the topic of saving money; by using the phrase "with

the development of economy", the learner is speaking in favour of the change of attitude, from that of saving money to that of spending money.

The high frequency and the unique formal and functional features of this chunk indicate obvious socio-cultural influences on language use. In the past decades, all sectors of the Chinese society, its economy, science and technology included, have witnessed unprecedented growth, and 'development' has been a catchword or hot topic. All this has found its way into learner interlanguage.

7. Conclusions and implications

In this paper, I have adopted Altenberg (1998) as a referential framework to investigate the phraseological features of Chinese learner spoken English. A large number of lexical sequences of various lengths and at different grammatical levels have been investigated and documented. By way of comparison, similarities and differences between the chunks in native data and in learner data have been discussed, in both quantitative and qualitative terms. Findings of the present study can be summarized around the following points.

Firstly, both Altenberg (1998) and the present study show a similar trend in data distribution across three structural types of chunk: full clauses, clause constituents and incomplete phrases, with clause constituents accounting for the largest proportion, followed by incomplete phrases, and then by full clauses. However, full clauses accounted for a much smaller proportion of the occurrences of all chunks in the learner corpus than is the case in Altenberg's (1998) data (4.6 percent as against ten percent). This indicates that the learners are rather weak in employing clause-level ready-made chunks to realize pragmatic meanings. It has been argued (Cowie 1988) that their inability to select these pragmatically specialized chunks for pragmatic functions in given situations is attributable to their limited language proficiency. Such an inability may also partly account for their non-nativelikeness and dysfluency in speaking.

Secondly, in terms of functional categories of the chunks, Chinese learners in comparison with native speakers, used a great deal more recurrent chunks basically associated with the making of propositions, while using far fewer (and much less frequently) chunks typically fulfilling pragmatic functions. More specifically, within clause constituents, learners frequently used unique categories of chunks, such as rhemes, nominal expressions and purposive expressions, which are basically propositional chunks. In contrast, chunks typically fulfilling pragmatic functions are obviously far less recurrent in COLSEC than is the case for Altenberg (1998). For example, functional categories, such as intensifiers and vagueness tags, have been greatly under-represented in the learner data. I have therefore argued that Chinese learner English speech is strongly biased toward the making of propositions, while the fundamental pragmatic qualities of discourse, such as cooperativeness and politeness, are basically neglected, which seems to be a common problem in second language speech.

Thirdly, learners have developed their own characteristic ways of realizing evaluative meanings. Take the sequence *it is very ADJ*, for instance. This occurred frequently in COLSEC, typically with an adjective such as *good*, *important*, *convenient*, *bad*, *interesting*, *helpful*, *useful* or *necessary*, and was characteristically used to comment on and evaluate things. As well as this, socio-cultural factors seem to have exerted a strong impact on language use. Owing to the fact that China has, as said previously, been undergoing rapid development, the expression *with the development of (society*, *economy*, *science and technology*) has taken on a particular semantic prosodic meaning and prominence, becoming a favourite device for the learners to air positive views and evaluations on social phenomena and to open discussions on a topic.

The findings of the present study have potentially useful implications for SLA studies and second language pedagogy. Both native speakers and foreign language learners rely heavily on particular phraseological sequences, whether at the sentential, clausal or phrasal level, for constructing discourse. But, clearly, the chunks typically used by learners differ dramatically from those of native speakers for given meanings and functions. Thus, to what extent and in what characteristic ways learner chunks display unique idiosyncratic formal and functional features merits more thorough investigation in SLA studies, and can shed new light on the nature and feature of second language development. Pedagogically, the study points to the necessity of giving more weight to the pragmatics of language, particularly in the teaching of conversation. It is desirable that more emphasis be attached to the appropriate use of functional or evaluative chunks in teaching English conversation, so that learner speech can sound more interactively competent, cooperative and polite.

Finally, it must be admitted that some of the idiosyncratic features of the learner phraseology reported on in the present study are partly reflective of the nature of COLSEC. Caution must be exercised when interpreting the reported data, to avoid hasty generalizations. This notwithstanding, the present study has illustrated that many characteristic features of learner phraseology in English speech can be uncovered and generalized by exploring and scrutinizing the data of COLSEC.

Notes

1 I would like to express my sincere thanks to Professor Antoinette Renouf, who kindly reviewed the draft of this paper carefully, with valuable comments on it and practical suggestions for its revision. I am also grateful to my colleagues in the COLSEC team, particularly Professors Li Wenzhong and Pu Jianzhong who devoted their valuable time, energy and wisdom to the construction of the corpus, evidence from which provided a firm basis upon which this study could be carried out.

2 More precisely, at COBUILD, Sinclair and his colleagues adopted a methodology whereby the corpus serves as an empirical basis from which lexicographers extract their data and detect linguistic phenomena without prior assumptions and expectations. But, it is Tognini-Bonelli who first defines this methodology as the corpus-driven approach (cf. Tognini-Bonelli 2001).

3 COLSEC sampled the CET test episodes within the period 2000 to 2004. A total of 302 episodes are covered in the corpus. I express my sincere thanks to the National College English Testing Committee for letting me use the material.

4 For details of sampling and annotation of COLSEC, refer to Wei, Li and Pu (2007).

5 Cf. Altenberg (1998): full clauses: ten percent; clause constituents: 76 percent; incomplete phrases: 14 percent.

6 According to Chinese philosophy, the heaven represents *yang*, while the earth represents *yin*; the mountain represents *yang* while the river represents *yin*; man represents *yang* while woman represents *yin*, etc.

7 I am grateful to Professor Antoinette Renouf for bringing this important point to my attention. As a matter of fact, discussion in the present section grew out of talks with her.

8 In table 7, the data for *I think (that)* includes: *I think that*: 231 times; *I think maybe*: 78 times; *I also think*: 45 times; *I think it's*: 651 times; *I think this*: 207 times; *I think they*: 163 times; *I think that's*: 132 times.

9 This sequence is arguable, since, by definition, a stem does not include any following subject or complement. But it could also be thought of as a stem on the grounds that the word *very* is not a substantial complement. Altenberg (1998: 114) put the sequence *what is the* into the category of interrogative stem.

10 COLSEC covers 39 conversational topics, which include 'job-hunting', 'saving money', 'caring for others', etc.

11 In Altenberg (1998: 117), frequencies of vagueness tags are: *and so on*: 47; *or something like that*: 16; *and all that*: 14; *and things like that*: 14; *something like that*: 13; *sort of thing*: 11; *that sort of thing*: 10. Considering the size of the *London-Lund Corpus*, 50,000 words in total, counterparts in COLSEC are significantly under-used except for *and so on*, which is significantly overused.

12 *With the development of* did not occur in the *London-Lund Corpus*. It occurred only four times in the spoken component of the *British National Corpus*.

References

Aijmer, K. (2002), *English Discourse Particles*. Amsterdam: John Benjamins.

Altenberg, B. (1998), 'On the phraseology of spoken English: the evidence of recurrent word-combinations', in: A.P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. 101-122.

Biber, D. (2004), 'If you look at...: lexical bundles in university teaching and textbooks', *Applied Linguistics*, 25(3): 371-405.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman Grammar of Spoken and Written English*. London: Pearson Education.

Bolinger, D. (1976), 'Meaning and memory', *Forum Linguisticum*, 1(1): 1-14.

Channell, J. (1994), *Vague Language*. Oxford: Oxford University Press.

Cowie, A. (1988), 'Stable and creative aspects of vocabulary use', in: R. Carter and M. McCarthy (eds.) *Vocabulary and Language Teaching*. London: Longman. 126-139.

De Cock, S. (1998), 'A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English', *International Journal of Corpus Linguistics*, 3(1): 59-80.

De Cock, S. (2000), 'Repetitive phrasal chunkiness and advanced EFL speech and writing', in: C. Mair and M. Hundt (eds.) *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*, Freiburg im Breisgau 1999. Amsterdam: Rodopi. 51-68.

De Cock, S., S. Granger, G. Leech and T. McEnery (1998), 'An automated approach to the phrasicon of EFL learners', in: S. Granger (ed.) *Learner English on Computer*. London and New York: Addison Wesley Longman. 67-79.

Dines, E.R. (1980), 'Variation in discourse – "and stuff like that"', *Language in Society*, 9: 13-31.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Adward Arnold.

Louw, B. (1993), 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in: M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins. 157-176.

Nattinger, J.R. (1988), 'Some current trends in vocabulary teaching', in: M. McCarthy and R. Carter (eds.) *Vocabulary in Language Teaching*. London: Longman. 63-82.

Pawley, A. and F. H. Syder, H. (1983), 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency', in: J. Richard and R. Schmidt (eds.) *Language and Communication*. New York: Longman. 191-225.
Pu, J.Z. (2003), 'Collocations, colligations and chunks in the teaching of English lexis', *Foreign Language Teaching and Research*, 6: 438-445.
Renouf, A. and J. Sinclair, (1991), 'Collocational frameworks in English', in: K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman. 128-143.
Sinclair, J. (1987a), 'The nature of the evidence', in: J. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins. 150-159.
Sinclair, J. (1987b), 'Collocation: a progress report', in: R. Steele and T. Threadgold (eds.) *Language Topics: Essays in Honour of Michael Halliday*. Amsterdam: John Benjamins. 319-331.
Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Tognini-Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.
Tsui, A.B.M. (1994), *English Conversation*. Oxford: Oxford University Press.
Wei, N.X. (2002), *Towards Defining Collocations*. Shanghai: Shanghai Jiao Tong University Press.
Wei, N.X. (2004), 'A preliminary study of college learners spoken English corpus', *Modern Foreign Languages*, 2: 130-140.
Wei, N.X., W.Z. Li and J.Z. Pu (2007), 'Design principles and annotation methods of the COLSEC corpus', *Contemporary Linguistics*, 3: 235-246.
Yang, H.Z. (1999), 'Design principles of the spoken English test of CET', *Foreign Languages World*, 3: 48-57.

Appendix

Concordances for *I see*

1	the bicycle </sp3> <sp1> Now,	I see.	I see all [Mll] you mean.
2	</interlocutor> <s3> Oh,	I see.	Er...I think, I don't
3	</interlocutor> <sp3> Oh.	I see.	I see. Erm I think in
4	</interlocutor> <sp2> Oh,	I see,	I see. Then I turn to the
5	much poorer </sp3> <sp2> Oh,	I see.	</sp2> <sp3> Yes. </sp3>
6	</interlocutor> <sp2> Ok.	I see,	the pictures describe the
7	</interlocutor> <sp2> Ok,	I see.	Er I think now people's
8	to them more. </sp3> <sp1> I see,		you opinion is just er as
9	to take a walk? </sp2> <sp1> I see.		I hope so. But you know
10	is ready, now. </sp1> <sp2> I see,		cigarette is bad for my

11 will you go abroad? </sp2> <sp3> I see. </sp3> <sp2> I can't
12 </interrupted> <sp3> I see. </sp3> <interlocutor> Ok
13 </interlocutor> <sp3> Yeah, I see I see. </sp3>
14 </interlocutor> <sp1> Yes, I see. </sp1> <interlocutor> Now
15 </interlocutor> <sp1> Yes, I see. Some people around me mm

Concordances for *It it's very*

1 with people with strangers, it's very bad, I think. So
2 I don't think so, so, I think it's very bad. Because we can
3 bus [Wu-er] mm go to school. and it's very convenient to our
4 eave this room, I would think [Mk] it's very essential because I
5 [W2n-ng] and sometimes I think it's very good to have one
6 in so in some working experience. It's very good for your
7 This is this is a well-paid job. It's very great. Yes. But
8 and [Pd-er] er...you know it's very helpful and it can
9 different people so I really think it's very helpful for me, and
10 to feel very [Wv-w] happy I think it's very important for your
11 I think er people with competition, it's very important because
12 China to another so I think [Wth-s] it's very interesting. Do you
13 continue their study so I think it's very necessary. Thank
14 interested in marketing. I think it's very useful, especially
15 ball is drop down. And I think it's very terrible. Em so I

Frequency of nominalization in Early Modern English medical writing

Jukka Tyrkkö and Turo Hiltunen[1]

University of Helsinki

Abstract

Nominalization has been noted as one of the fundamental properties of scientific English. It allows the dense packing of complex ideas into elements of clause structure, the addition of modifiers and qualifiers, and the backgrounding and foregrounding of information in the discourse. The device of nominalization in scientific prose is said to originate with Newton's writings in the seventeenth century. This paper presents the results of an examination of a 1.7 million-word corpus of Early Modern English medical texts with the aim of determining whether a significant increase can be observed in the frequency of nominalizations between 1500 and 1700. The identification of nominalizations in the corpus is operationalized using a defined set of nominal suffixes. Our results suggest that nominalizations do not first emerge during the period of Empiricism in the second part of the sixteenth century, but that, at least in medical writing, a quantitative increase in nominalizations is observable even earlier: our data shows an increase in the frequency of nominalizations throughout the period under investigation. However, there is also a great deal of variation between individual texts throughout the 200-year period: many texts from the sixteenth century are entirely comparable to seventeenth-century texts in terms of nominalization density.

1. Introduction

The basic definition of 'nominalization' is that a noun – or, derivatively, a nominal structure – is used in the place of a verbal expression (see examples (1a) and (1b), respectively, italics added in all examples).

(1a) Not satisfied with my *Observations* thereon, I continued
my Endeavours to discover the true Texture of Bones;
(Leeuwenhoek, *The Philosophical Transactions*, 1693 (198): 838)

(1b) For proof hereof, I *observ'd*, that Men would lie all
night, and sleep on the Sands without hurt.
(Stubbes, *The Philosophical Transactions*, 1668 (3): 701)

As a noun, a process can function as the head of a noun phrase and, consequently, as the subject or object of a sentence. This makes adjectival modification possible

as well as the use of an extended range of cohesive devices such as pronominal reference, and helps to focus thematic attention on a process by changing its semantic role. Perhaps more important still, nominalization allows the packing of large amounts of information into single lexical items. Such nouns often begin to acquire additional and expanded meanings beyond those immediately traceable to the corresponding verb.[2] All this makes nominalizations a useful discursive strategy in various genres of writing where processes themselves are the object of discussion, such as scientific writing (see e.g. Halliday and Martin 1993; Myers 1990; Ventola 1996). According to Halliday (1988), nominalizing began to increase notably toward the end of the seventeenth century and became a prominent part of the emerging scientific register during the early eighteenth century. Discussing Newton's *Treatise on Opticks* as an example, Halliday argues that the reificative function of nominalizing reflected a fundamental change in the way the world of natural phenomena was conceptualized in natural sciences. More recent studies by Atkinson (1999) and Banks (2005b) have both corroborated Halliday's findings and posited additional explanations. We shall examine the latter in section 5.

This paper presents the results of an examination of a 1.7 million-word corpus of Early Modern medical texts with the aim of determining whether a significant increase can be observed in the frequency of nominalizations. Our study is different from previous analyses in several important respects. Firstly, it not only looks at the first decades of the period of empiricism in England, but also takes into account developments preceding that period. Moreover, unlike many previous studies, our analysis is not based on texts of natural philosophy, but focuses on medicine, a discipline with the longest tradition in the vernacular. The period under investigation is an important one in the history of medicine as it coincides both with general changes in the thought-style of natural sciences and with major field-specific changes and innovations. The Early Modern period saw the decline of the Galenic medicine that was based on Aristotelian natural philosophy, and the rise of empiricist 'new science' (Wear 1995). From this perspective, we view nominalization as one potential linguistic correlate of such changes in the paradigm of medical science.

Complementing earlier studies, our corpus-based approach provides quantitative, statistically testable evidence on the use of nominalization in printed medical texts between 1500 and 1700. Methodologically, the study explores the possibility of analyzing high frequency phenomena in an unannotated corpus. Concurring with Heyvaert (2003a: 4), who points out that "nominalizations do not always have discernible structural components for each function which they can realize", we note that the detection of such structures from an unannotated historical corpus would be prohibitively time consuming. Accordingly, we leave aside nominalizations realized by means of clausal structures such as *that*-clauses, infinitive structures, etc. used in the place of nouns (cf. Schmid 2000: 65; Vendler 1967: 122-140). We likewise exclude general nouns as defined in Halliday and Hasan (1976). Instead, we concentrate exclusively on word level nominalizations in a manner similar to Biber (1988: 227), claiming that the volume of occurrence

of suffix-marked nominalizations, in our case over 17,000 in total, is sufficient for the detection of developmental trends.

Our results confirm the trends identified in earlier work, but suggest that, at least in medical writing, a quantitative increase in nominalizations is observable even before the period of Empiricism: our data shows a steady increase in the frequency of nominalizations throughout the period from 1500 to 1700. At the same time, we note that there is a great deal of variation between individual texts throughout the 200-year period, with many texts from the sixteenth century entirely comparable to seventeenth-century texts in terms of nominalization density.

2. Defining and identifying nominalizations for corpus analysis

This study focuses on action nominals, that is, nouns which derive from verbs and which refer to an action or, primarily as a result of semantic expansion, a set of actions. From a methodological point of view, the major issue to tackle is how to retrieve these with reasonable precision from a corpus. There is a wealth of research on nominalization covering aspects of form, meaning and use. However, this body of research is not immediately useful for our purposes, as different studies have either used a different definition of nominalization or operationalized it differently for corpus searches. The inconsistencies often appear to arise out of the fact that the term nominalization can be used in reference to both the word formation processes with which new nouns are created and to a systemic functional concept where a noun or nominal structure is taken to stand in the place of a verb (Heyvaert 2003b).

As has been done in several previous corpus-based studies, the identification of nominalization in this study is operationalized using a defined set of nominal suffixes. This 'suffix method' of identifying nominalizations has previously been applied in multi-dimensional studies, starting with Biber (1988). While the method is relatively straightforward and has proven to be robust, it has certain limitations. The alternative would be to adopt a purely systemic functional definition, which would be more inclusive in its coverage. However, the systemic functional approach is not practicable in a corpus study of this magnitude, and for this reason we have opted for the suffix method, taking great care in formulating the operative definition for the category nominalization in an appropriate fashion.

There are two major issues to take into account when it comes to applying the suffix method. The first of these is the fact that it is based on a formal definition of nominalization, and therefore does not cover all the expressions that would be considered nominalizations under the functional definition. For instance, various clausal structures such as infinitive clauses, as well as nominalization involving a lexical change, are ignored by such a formal definition. Nouns derived from verbs by means of conversion are similarly left out. Accordingly, a definition based on a closed set of suffixes would not treat the noun *use* as a nominalization of the verb *use*, only nouns like *using* and *usage.* To retrieve all such occurrences would involve going through thousands of forms in the word list that

could potentially be either verbs or nouns, many of them high-frequency words like *use*, and analyzing them individually. Such a task would not have been feasible with a 1.7 million-word unannotated corpus.

Another problem is the fact that despite the availability of data on the use of nominalization in various kinds of scientific texts, previous studies are not necessarily commensurate due to methodological differences. Individual scholars have opted to use considerably different sets of suffixes, and, to complicate matters further, varying levels of other criteria for inclusion or exclusion of particular types of nouns. As a result, although the results of the various studies appear to be in agreement on major issues, they are not immediately comparable.

The three nominal word formation processes are derivation, compounding, and neologizing;[3] only derivation is of interest here. Although nouns can be derived from other nouns – such nouns can be called denominal nouns – and adjectives, the prototypical process involves the addition of a nominal suffix to a verb base.[4] For the Early Modern period, Nevalainen (1999: 391-400) gives the following deverbal noun suffixes:

> *-ation*, *-ment*, *-ance/ence*, *-al*, *-ing*, *-ure*, *-age*, *-ant/-ent*, *-ard*, *-ee*, *-er*

and the following deadjectival suffixes:

> *-acy*, *-ancy/ency*, *-by*, *-ity*, *-ness*, *-ton.*

Identifying nominalizations as one of the several features of dimension 3, "situation dependent vs. explicit reference", Biber (1988: 227) operationalizes nominalizations as "all words ending in *-tion*, *-ment*, *-ness* or *-ity* (plus plural forms)". Looking at Nevalainen's lists above, we see that Biber includes two suffixes each marking deverbal and deadjectival derivation. Atkinson (1999: 138) follows Biber's lead in his diachronic multidimensional study of the *Philosophical Transactions of the Royal Society*, but expands the list of nominal suffixes to include the following, described as markers of "abstract nouns":

> *-tion*, *-ment*, *-ity*, *-ess*, *-al*, *-dom*, *-tee*, *-er*, *-or*, *-ence*, *-ance*, *-ant*, *-lty*, *-ure*, *-ery*, *-cy.*

A comparison shows that Atkinson includes five suffixes not on Nevalainen's list (*-ity*, *-dom*, *-ess*, *-lty* and *-ery*) and excludes two (*-ing* and *-ard*).[5] Atkinson does not discuss the reasoning behind the composition of the list, but it seems likely that at least the exclusion of *-ing* may be due to the fact that, despite the high overall number of occurrences of the suffix, only a minority are nominalizations. Like Biber, Atkinson does not make it clear whether or not any close examination of the retrieved items was conducted. Seeing as the unfiltered results of a search for the suffix *-tion*, for example, can easily yield results such as *motion*, *question*, and *nation*, none of which would be easy to class as nominalizations, this raises a potential point of contention. Naturally, the issue would be made even more

complicated if the results obtained were to be examined in comparison to, or in contrast with, those of functionally grounded studies. It goes without saying that both Biber and Atkinson use lists as a workable way of deriving nominalizations out of a corpus, and that neither suggests that the suffixes listed could be considered a comprehensive account of nominalization.[6]

Under the functional paradigm, the guiding principle is the relationship between the noun in question and the semantic function it stands for. This approach is based on Halliday's (1988) concept of 'grammatical metaphor',[7] which refers to cases where semantic 'processes' are realized incongruently as nouns instead of verbs, which would be the congruent realization. By this definition, nominalization is limited to deverbal and deadjectival nouns. Banks (2005b) accordingly draws a distinction between a nominalization and a deverbal noun as follows: while deverbal nouns also include nouns functioning as agents, instruments or results of processes, nouns functioning as grammatical metaphors must refer to the process itself.[8] From the perspective of studying nominalization as a feature of scientific writing, the functional definition is without question more relevant than the more limited conceptualization based on nominal suffixes. However, systemic functional analysis would mandate the close examination of each noun in context to discover the corresponding agnate, or indeed a network of all agnate structures, as argued by Heyvaert (2003b: 68). It is clear that such a task is virtually impossible to operationalize in a quantitative corpus study. Existing corpus studies following the functionalist approach, e.g. Banks (2003, 2005a, 2005b), have solved the problem by using so called 'mini-corpora' and analyzing them exhaustively. However, such small datasets allow for only cautious generalizations, particularly for the purposes of longer diachronic studies.

Having access to a larger corpus, we opted for a different approach. Given the frequency at which suffix marked nouns occur in writing, we would argue that the method is a sufficiently reliable indicator of the degree to which nominalizations are used overall. The quantitative evidence this approach provides complements previous scholarship largely conducted from the qualitative perspective.

2.1 Method followed in this study

From the different approaches discussed above, it is apparent that no single, widely adopted standard exists for the morphological identification of nominalizations. We therefore have to find a balance between accomplishing the task of deriving accurately identified high-volume data and maintaining practical feasibility. Aiming to produce comparable data, we took as our starting point the suffix list used by Atkinson (1999). For practical reasons, three of these suffixes were removed. Firstly, because the corpus is not part-of-speech tagged, the inclusion of the most productive suffixes *-ing* and *-er* was not practicable.[9] The discarding of agentive suffixes *-er* and *-ee* was further supported by the functional paradigm as defined by Banks (2005b). As for *-ing*, the sheer volume of tokens, arising from the similarity with the present active participle or *-ing* form, imposes a practical limit. After removing these, we were left with a list of 13 suffixes (see below).

Because our primary research question derives from an interest in examining the development of scientific thought-styles, the premise is that thought-style shifts are most conspicuous in contexts where, at least conceivably, the author would have had recourse to a more or less synonymous active verb. Continuing this line of reasoning, we take the availability of a corresponding verb in contemporary lexicon as a sufficient indication that the writer would have had access to a congruent agnate verb form that could have been used instead. To operationalize the selection process of the nominalizations to be included, we would only include nouns which fulfilled all three of the following criteria. To be included in the study, a noun had to be:

- overtly marked with a deverbal or deadjectival nominal suffix from a preset list: *-ion*, *-ment*, *-ity*, *-ness*, *-al*, *-dom*, *-ence*, *-ance*, *-ant*, *-lty*, *-ure*, *-ery*, *-cy*;
- derived from a verb available in contemporary lexicon (based on the datings in the *Oxford English Dictionary*);
- in the case of deverbal nominalizations, reiterable in the context with a variant expression where the semantic content is encoded in a verb.

We then searched for all instances of words containing one of these suffixes and their plural varieties, in all spelling variants. Following that, we discarded instances which were not nouns or verbally reiterable by the contemporary writer. The *Oxford English Dictionary* was consulted only in uncertain cases, and the reiterability was discussed on a case-by-case basis when deemed necessary. With some words, the application of these criteria proved to be problematic. For example, should *description* be included, seeing as no corresponding verb was available in contemporary lexicon to function as a base form (e.g. *descript)? Because we are primarily interested in the frequency of nominalizations in scientific writing, rather than the word-formational aspects of individual lexemes, we argued for inclusion whenever the relationship between the nominalization and a contemporaneously available verb was evident.

The frequency of nominalizations in a text is taken as an indicator of their importance as a rhetorical resource in medical writing. To compare the density of nominalizations in texts of different lengths, frequencies were normalized to 1,000 words of running texts. Averages of normalized frequencies for different categories of writing were counted and compared, as well as those of four 50-year periods between 1500 and 1700. The results were tested for significance using the non-parametric ANOVA test.

3. The EMEMT corpus

The *Early Modern English Medical Texts* (EMEMT) corpus (forthcoming) used in the study has been compiled by members of the Scientific Thought-styles project at the University of Helsinki. A corpus of 1,85 million words,[10] EMEMT is composed of extracts of printed medical texts published between 1500 and 1700 (in the case of category 6, extracts from the *Philosophical Transactions*, between

1665 and 1700). The texts are divided into six categories and an appendix, on the basis of extralinguistic parameters (see figure 1).[11]

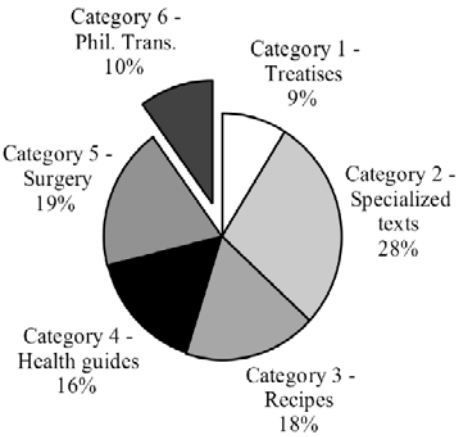

Figure 1: EMEMT word count distribution by category

The categorization of EMEMT reflects the textual reality of contemporary medical science and can be used as a starting point for conjecturing possible reasons behind linguistic phenomena. Although EMEMT is weighted somewhat toward the more prestigious texts, each category also contains texts by authors who had no official qualifications. Any such generalizations must, however, be handled with caution. In broad terms, texts belonging to categories 1, 2, 5, and 6 represent the learned end of the spectrum, while those in categories 3 and 4 are more likely to have been written by writers without a formal education. Education is naturally a fluid and context-sensitive concept, and we will do well to remember that in sixteenth- and seventeenth-century England medical practitioners came from a variety of backgrounds (see e.g. Wear 1998 or French 2003). The number of university-trained physicians was very low in England, but it does not necessarily follow that a writer without university qualifications could not read Latin or be well-versed in contemporary medical practice.

As noted, the timeline of the texts studied here predates Halliday (1988) and Banks (2005b) and overlaps only just with Atkinson (1999). While this positions us well to examine the period preceding the Scientific Revolution of the eighteenth century, it must be noted that EMEMT consists exclusively of medical texts and therefore does not sit in a direct line of continuum with the natural sciences texts examined by Halliday (1988) and Banks (2005b). We would nevertheless maintain that the makeup of the learned community in late Renaissance England and the closely connected nature of different genres of learned writing make findings from the medical corpus relevant to evaluating the prevalence of nominalizations also in other learned genres.

4. Findings

Following the method outlined above, 17,163 tokens of nominalization were derived from the corpus. An overwhelming abundance of these were formed with the suffix *-ion*, which alone accounts for 86 percent of the total (see figure 2). Along with conversion, *-ion* and *-ment* suffixation are noted by Bauer (2001: 180) as particularly successful processes of nominalization in the history of English, and EMEMT confirms this to be the case.

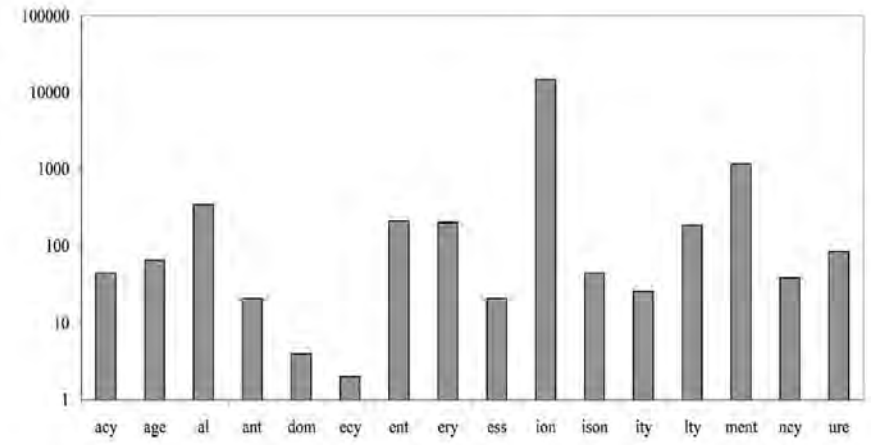

Figure 2: Distribution of nominalization suffixes (absolute frequency, logarithmic scale)

To discover overall developmental trends, we examine the corpus in four sections of 50 years each. We prefer this structuring of the timeline over the many possible alternatives based on, for example, historical events, as it is difficult to evaluate how and to what extent they influence linguistic reality. Having said that, the periodization followed here should also serve relatively well in illustrating developments in medical writing. The first section roughly corresponds with the last decades during which scholastic medicine still prevailed, the middle two with times of increasing empiricism, and the last section with the beginnings of the era sometimes called the Age of Reason. The last section notably also includes the collection of articles from the *Philosophical Transactions of the Royal Society*, first published in 1665.

We begin the analysis of nominalization frequency by collapsing the texts of categories 1 through 5 into one and comparing the resulting four slices of the timeline with each other (figure 3).

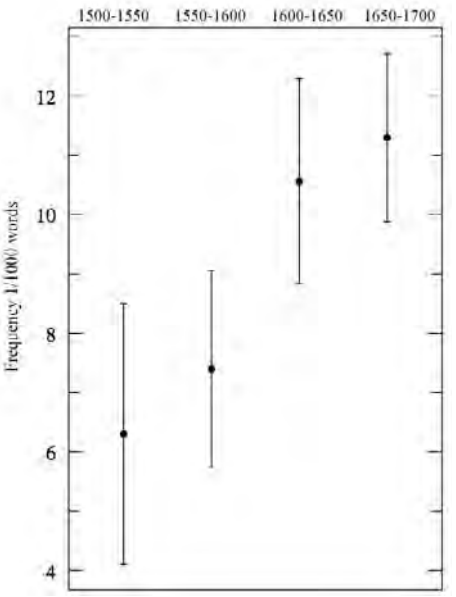

Figure 3: Pooled averages with 95 percent confidence intervals (frequency 1/1000 words)

A statistically significant (p=0.0001 by the non-parametric ANOVA test) increase in the use of nominalization is observed. Interestingly, the evidence also makes it clear that the trend did not appear toward the end of the seventeenth century but on the contrary it seems that, at least in medical writing, the use of nominalization has increased since the early sixteenth century. When the categories are examined individually in each of the four sections, however, the picture becomes somewhat more complicated. Leaving aside the 'treatises' category for the moment,[12] we move to analyzing categories 2, 3, 4 and 5 across the timeline (figure 4, below).

Statistical testing shows that the differences between the categories are borderline significant until the last 50-year section, at which point the significance becomes clear.[13] However, it is equally clear that this result applies only to the 'recipes' category (second line from the left in the figures above), while the differences between the other three categories are not significant. Two observations in particular can be made. First, that recipe collections differed from the other categories in some consistent and systematic fashion, and second, that the increase in nominalization in medical writing was not driven by developments in one particular subgenre, but rather took place across the board, more or less at the same rate.

The first observation is relatively easy to explain with what we know of medical recipes. The instructional nature of recipes, enacted through imperative constructions (see example (2) below), naturally lends itself less to the use of nominalizations (cf. Mäkinen 2006: 87-94).

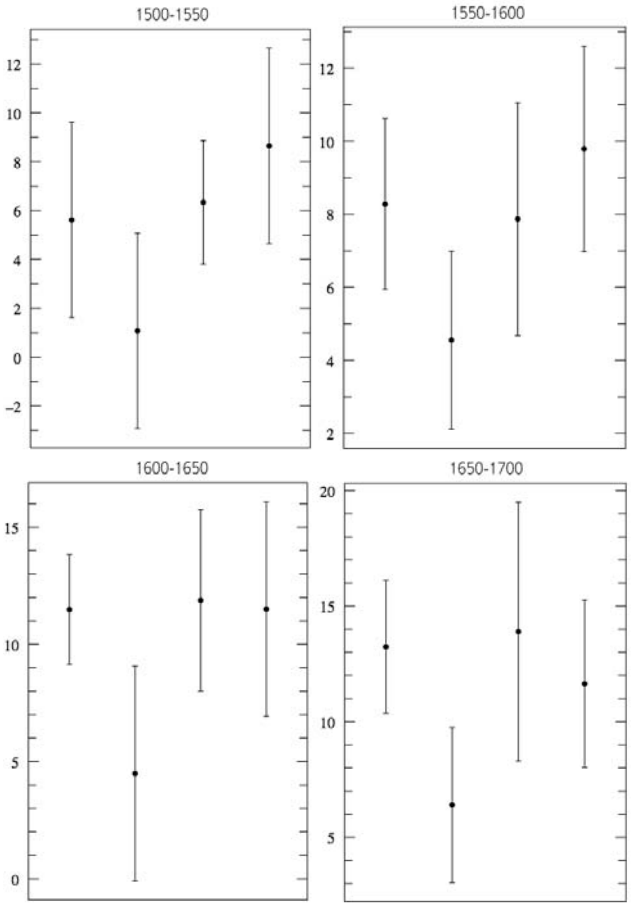

Figure 4: Nominalizations in categories 2, 3, 4 and 5 in 50-year sections of EMEMT

(2) [{ The fyfte water. }]
The fyfte water is suche that with it you may do
many merueylous thynges. Take Lymell of syluer
Golde Latyn Coper Iron Stele & lede. Also take
Lytarge of golde & syluer / & take Camemell & Columbyne
and stepe all togyder in the vryne of a man
chylde that is made by a daye and a nyght. The seconde
day in whyte wyne. The thyrde day in y~ ioyce
of fenell. The fourthe day in y~ whyte of egges. The
fyfte day in womans mylke that noryssheth a man
chylde. The syxte day in reed wyne.
(anon., *The Treasure of pore men*, 1526)

The second observation is a much more far-reaching one. As noted in section 3, the texts of the different categories were produced by diverse communities of medical practitioners; for example, surgeons, physicians and apothecaries formed separate and at times antagonistic factions. Also, on the basis of earlier research on Late Medieval medical writing as well as preliminary findings using EMEMT,[14] we would expect that such seemingly diverse subgenres as 'specialized medical texts', 'health guides' and 'surgical treatises' would show more marked differences in the rate at which they accommodate a new linguistic feature – particularly when that feature is consistently described in modern scholarship as a feature of formal, scientific style. However, our data does not appear to give support to this hypothesis. The lack of statistically significant difference between categories 2, 4 and 5 leads us to surmise that an increase in the use of nominalization took place universally in factual discourse, rather than only in the highest echelons of scientific writing. Naturally, this does not dispute the fact that nominalization came to mark scientific writing in general, particularly from the early eighteenth century onwards.

The collection of medical articles from the *Philosophical Transactions* included in EMEMT has been left to the side until now. Known as the first English scientific journal – and second only to the French *Journal des sçavans*[15] – the *Philosophical Transactions* is widely credited as having spearheaded the rise of modern-style scientific writing (Atkinson 1999). The articles are a mix of letters to the Royal Society, scientific reports of experiment and observations, and review articles. On the whole, the texts are characterized by a decidedly empirical flavor with references to old authorities notably scarce compared to contemporary medical books, and much of the writing reflects the personal observations and experiments of contributors. This mixture of manifestly modernistic approach to writing on the one hand, and of first-person accounts on the other, makes it difficult to predict how the journal might reflect the rise of nominalization. The observed frequency of nominalizations in the *Philosophical Transactions* is 15.9/1000 words. In comparison to the other categories, the *Philosophical Transactions* thus appear well in line with contemporary medical writing. The figure is very close to Atkinson's (1999: 122), who gives 17.1/1000 for the *Philosophical Transactions* in 1725.[16] As suggested above, comparisons with Atkinson's figures are made with two caveats, namely that the list of nominal suffixes differs from ours by including *-er* and, more importantly, that Atkinson's selection of articles comprises a wider topical range than category 6 in EMEMT. However, Atkinson's figure can be taken as further testimony of the fact that, when it comes to the use of nominalization, the *Philosophical Transactions* did not lead the way.

To further compare the results with previous research, we applied the analysis to chapter 3 of Newton's *Opticks*, the text held up by Halliday and Banks as an example of the nominalizing style of scientific register.[17] Following the same principles used in the analysis of the corpus, we found a frequency of 36.9/1000 words. Although higher than anything found in the medical texts, the frequency is clearly still within range of EMEMT figures for the end of the seventeenth century. Several EMEMT texts come quite close to Newton. For exam-

ple, Gideon Harvey's *Morbus Anglicus* (1666) and Maynwaringe's *Methods and Means* (1683) show frequencies of 32.03 and 28.06, respectively. Example (3) shows the conceptually dense style of Maynwaringe:

(3) And therefore by a diligent inquisition,
and curious speculation into the works of
Nature, you may as much admire the manner
of preservation, government, order, weight,
and measure, regular vicissitudes, alternations
and successions; as the excellency and contrivance
of the things themselves in their
creation and generation.
(Maynwaringe, *The method and means of enjoying health, vigour, and long life*, 1683: 18)

Although Maynwaringe was a prominent physician, the health guide was written for the literate middling classes, and was not therefore a scientific treatise. The excerpt shows, however, that the author makes no concessions to readerly comprehension, employing instead a range of highly specialized terms and a wealth of nominalizations.

As noted already, examples of authors using nominalization very frequently are extant already from the sixteenth century. A good example is Timothy Bright (1549-1615). A widely travelled physician and an early exponent of shorthand writing, Bright wrote on a number of topics ranging from religious treatises to cryptography. What makes Bright particularly interesting is that despite his education and obvious language skills, he wrote in English as well as in Latin. His book on mental illness, the *Treatise of Melancholie* (1586) was an important influence on Burton's *Anatomy of Melancholy* (1621) and may have been used as background reading for *Hamlet* (see O'Sullivan 1926). Importantly for the present study, it was written almost exactly a hundred years before Newton's *Opticks*.[18] In the 10,107-word extract in the corpus, Bright reaches a nominalization frequency of 21.37/1000 words, easily on par with late seventeenth-century texts and almost three times the average for the period. The effect of such a frequency is decidedly modern, as seen in the following extract:[19]

(4) Whereby it should
seeme, that poyson it self, where a nature fitteth,
therewith may be matter of holesome *nourishment*,
for the *satisfying* of which *obiection*, we
are to consider euerie parte of that we take for
nourishment, is not alimentall but parte *excrement*,
and that the greatest parte, as it appeareth
by so manie *alterations*, and *purginges*,
which the foode suffereth, before it be receiued
of the partes of the bodie for proper *nourishment*.

> so therefore; these birds are not sustained
> with that which is poysonfull in their foode, but
> alter it first, and then passe it into superfluous
> *excrement*; their *substance* being vtterly voyde
> of the same, & so becometh vnto vs holesome:
> verie well: but how is their nature able to vanquish
> (Bright, *Treatise of Melancholy*, 1586: 20)

Bright's use of nominalizations has many similarities to that found in modern academic register. He makes use of adjectival premodification when talking about "familiar", "holesome" and "proper" nourishment. He nominalizes physiological processes ("nourishment", "purging"), their products ("excrement"), and properties ("substance"). Some of the structures Bright constructs are quite complex, as illustrated later on the same page (example (5)), where a semantically complex *that*-clause includes no less than three nominalizations.

(5) whereby it is manifest, that with natures
arte an apt matter of *producing* of *nourishment*
must needes meete for her *maintenance*.
(Bright, *Treatise of Melancholy*, 1586: 20)

4.1 Lexis and its distribution

As noted above (figure 4), a statistically significant difference (p=0.0004 by the non-parametric ANOVA test) is observed between recipe collections and the categories 2, 4, and 5. While the pooled averages for the treatises, specialized texts, health guides and surgical texts appear at first more or less similar, the picture is again more complicated when one delves deeper. So far the evidence would appear to support the view that the frequency of nominalizations increased over the two centuries, but it would be too hasty to therefore construe that nominalization would not have been available as a linguistic strategy much earlier. Indeed, if we look at the nominalizations text by text, it soon becomes apparent that a significant number of writers used a distinctly nominal style much earlier than is generally given. Ioannes De Vigo's *Most excellent workes of chirurgery* (1543), for example, shows a normalized frequency of 15.17/1000 words, which would not have been out of place a hundred years later. Likewise, Edward Iorden's gynecological treatise on the womb entitled *The Suffocation Of The Mother* (1603) shows an even higher frequency of 22.56/1000 (see example (6)).

(6) All these manner of wayes hath the Matrix by
consent to impart her offence vnto other parts. For
there wa~teth no corruption of humor, vapour, nor
euill qualitie, where this part is ill affected, to infect
other partes withall, there wantes no oportunitie
of conueyance or passage vnto any part, by

> reason of the large Vaynes, Arteries, and Nerues, which are deriued vnto it, with which it hath great affinitie and similitude of substance, besides the connexion it hath with the heart, liuer, braine, and backe.
> (Iorden, *A briefe discourse of a disease called the suffocation of the mother*, 1603, f. 8)

Given that the seventeenth century is identified in previous scholarship as the time when scientific nominalization began to emerge, there is a particular interest in examining it with closer attention. The wide variety observed in the employment of the nominal style continues into the seventeenth century (figure 5). It is particularly notable that not one of the five categories examined shows a tendency to uniformity, but rather they all appear to include texts from both ends of the nominalizing spectrum.

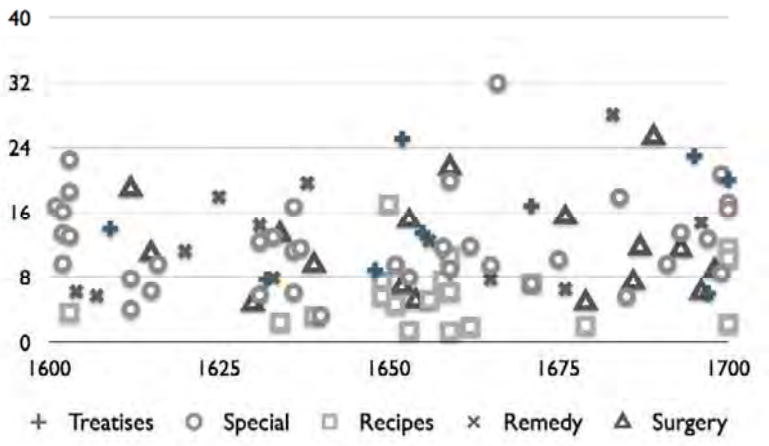

Figure 5: Nominalizations in individual texts in the seventeenth century (1/1000 words)

Another point of interest that pertains to the use of nominalization concerns the abundance of individual derivation types used in individual texts across the timeline. This measure differs in principle from the overall frequency of nominalizations in a text, the former being an indicator of lexical creativity and the latter of a nominalizing discursive style. By dividing the number of unique derivations with the word count and standardizing the value to 1,000 words, we get a ratio of unique nominalization types. Looking at the average of these ratios in 50-year sections makes it clear that the use of nominalization increased not only in terms of the sheer volume of tokens, but also in the range of types (see figure 6). Whereas a typical 10,000-word extract from the early sixteenth century has only

four or five types of nominalization, one of similar length from the end of the seventeenth century would have ten or more. The findings are statistically significant ($p<0.0001$ by non-parametric ANOVA).

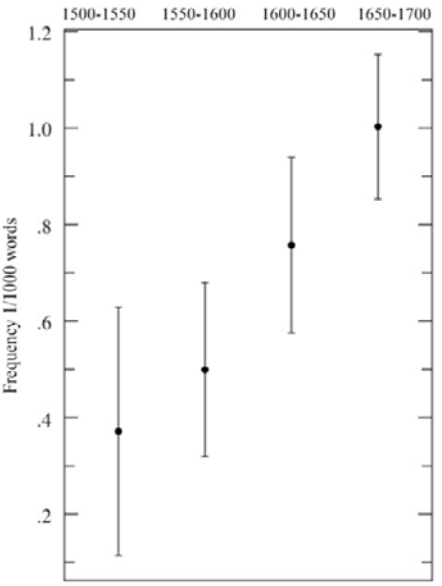

Figure 6: Average frequency of unique nominalization types for all categories combined (1/1000 words)

Interestingly, the writers who nominalize most flexibly tend to be some of the most noted writers of seventeenth-century medicine – a fact which may partly explain why the writers who came after them endeavored to affect a similar style. Nicholas Culpeper, William Harvey, and Everard Maynwaringe are some of those who show frequencies of over 2/1000 words. As was the case with overall use of nominalization, early examples are easy to find. John Ghesel's health guide entitled *Rules of Health* (1631), for example, exhibits a frequency of 2.21/1000 words, placing the author firmly among the most creative users of nominalization, despite the fact that his overall use of nominalizations is only slightly above average for the period at 14.36/1000 words.

By the same token, we can analyze each part of the timeline for the occurrence of individual types of derived nominals to get an approximate measure of usage. The majority of nominalizations occur across the timeline, that is, are attested in at least one text in each 50-year section. Although quite a high proportion show a clear preference for one end of the timeline or another, notably few appear exclusively in one century or another (table 1).[20] *Apostumation* (= abscess) is a good example, with only eight of the 91 occurrences in the corpus occurring in the seventeenth century. *Description*, on the other hand, occurs a total of 163

times, with only a single occurrence in the 1500-1550 period and 48 in the 1550-1600 period.

Table 1: Nominalizations exclusive to one end of the timeline

Nominalizations exclusive to the 16th century	Nominalizations exclusive to the 17th century
abomination, adulation, aggregation, ambulation, boogery, confession, embrocation, impostumation, plexion, repercussion	amputation, connexion, convulsion, defluxion, discovery, division, duration, evacuation, examination, improvement, injection, injunction, invention, management, palpitation, precipitation, prevention, revulsion, suppression

Most of the lexical appearances and disappearances can be explained by changes in the prevailing topics of medicine or more general changes in word sense. Newly discovered treatments, like injections, appear in the medical lexicon and naturally feature in written texts as new nominalizations. Borrowings and neologisms from Latin naturally play a major role, whether or not the term in question is a specifically medical one. Example (7) demonstrates this with "ambulations", used here in the general sense for 'walk', clearly with the intention of imparting a more learned air to the text.

(7) that those persons which feede vpon
grosse meates, and suche repast as bringeth
stronge nourishment, may vse vehementer
exercise and stronger *ambulations*,
as running, wrastling, hunting,
quoytes, handeball, and to be breife all
suche exercise as best agreeth with theyr
nature, which may brynge difficultie in
fetchyng the breath, augment heate and
humecte the skynne and exteriour parte
of the body, with a thin subtile sweate.
(Gratarolo, *A direction for the health of magistrates and studentes*, 1574)

When it comes to neologizing, medical writers were famous – or notorious, depending on the perspective – for the habit of creating new words (see Norri 1992). Nominalizations were one strategy available for accomplishing the task, and the EMEMT corpus provides plenty of evidence of such fledgling specialized terms.

5. Discussion: some tentative explanations

Before discussing the more pragmatically oriented explanations for increasing nominalization, we need to consider the possibility that the phenomenon could be accounted for by a general increase in the frequency of nouns. While we are unable to offer definitive answers due to the lack of part-of-speech annotation in EMEMT, it seems unlikely that the near doubling of nominalization from 1550-1600 to 1650-1700 could be accounted for simply by growing 'nouniness'. Nouniness is noted in literature as a feature of written discourse (Halliday 1988) and has been associated with formality (Brown and Levinson 1987: 207-209). Increasing use of nominalization could therefore be expected to go hand in hand with a more general developmental trend toward a standardized register of academic language. Did medical writers adopt a more nominal style for functional reasons – such as increased efficiency in the packing of information – or did the functional dimension only come about later, following a more stylistically driven desire to employ a newly popular linguistic strategy?

Banks (2005b) offers the influence of translation and the fact that the major academic authors of the late seventeenth and early eighteenth centuries wrote in both Latin and English as explanations. Examining Motte's (1729) English translation of Newton's *Principia Mathematica*, Banks argues that nominalized Latin forms were reproduced in the English translation. While this theory would appear convincing when it comes to the natural science texts examined by Banks, some doubts are raised when medical writing is added to the mix. Vernacular writing already began in medicine in the late fourteenth century (see Taavitsainen and Pahta 2004), and by the seventeenth century the vast majority of medical texts in England were written in English (Johns 2002: 283). Most early vernacular medical texts were translations from Latin. If Banks' claim is correct, it would seem reasonable to assume that such a strategy would have been adopted particularly eagerly in medical writing, where the influence of Latin, particularly on the lexicon, has been very significant (see Norri 1992). Thus, given the fact that the history of vernacular writing is considerably longer in medicine than in the natural sciences, we might expect nominalization to have emerged earlier than in the natural sciences. Our findings appear to refute this and instead suggest that nominalization did not become a major strategy in medical writing significantly earlier than in the natural sciences. On the other hand, the high frequency of nominalizations in some sixteenth-century original English compositions discussed in section 4.2 can be taken as an argument in favor of Banks' position. All the texts showing unusually high frequencies were written by Latinate authors.

Another explanation, suggested by Jucker,[21] is that seventeenth-century medical writers may have considered nominalizations a prestige feature and perhaps used them as something of a simulacrum for Latin, a language many of them no longer knew fluently (see McConchie 2002: 272-273). This explanation would be well supported by both linguistic and historical evidence. By omitting much of the information that is normally explicitly written out, highly nominal texts require more background knowledge. Importantly, such a style can signal a high

level of discursive competence, even if the subject matter of the text is familiar to the reader. Discussing texts featuring high frequencies of nominalizations, Ventola (1996: 185) comments that they "become enforcers of prestige and power, because the general public cannot understand the text". If the same holds true for Early Modern writing, which seems likely, there may have been motivations at play other than the purely functional one of packing information efficiently (cf. Halliday 1988: 149). As for the history of medicine, French (2003) argues convincingly that projecting an aura of learnedness was vital for both the social standing and financial interests of the Early Modern physician. All this helps explain the high incidence of nominalizations in category 4, 'health guides'. Naturally, health guides dealt with several topics which in themselves are conducive to nominalization, such as nourishment and digestion. At the same time, however, the primary audience for health guides were householders and the increasingly literate middling classes, and it is not unreasonable to think that nominalization may have been the perfect linguistic answer to the authors' need for putting across their own learning without using overly complicated terminology.

Halliday, Atkinson and Banks all emphasize how the increase in nominalization went hand in hand with a change in the underlying scientific thought-style. The argument goes that, by turning verbs of process into nouns, authors were able to refer to them as subjects or objects of arguments, and to modify them further or chain them together into increasingly complex noun chains. Reification like this was an integral part of the developing new register. Halliday and Martin (1993: 217) comment that "nominalized language is simply a symbol of literacy and thus education and thus power in our culture". The Early Modern texts provide us with a window into the time when such symbolism was first born, and as such allow us to shed some light on the process through which a linguistic strategy acquired a new social dimension. At least two questions are raised. First, to what extent was the increasing use of nominalization a matter of producing new words to meet a new need (whether intellectual or social), and second, to what extent were contemporary authors aware of using a new linguistic strategy? If we accept as a starting point the hypothesis that authors began to nominalize more around the turn of the eighteenth century, we can reasonably expect that this would have involved an increase in the use of both nominalizations established in the mental lexicon and neologisms using nominal suffixes (cf. Plag 2003). Our results (figure 6 and table 1) suggest that from the beginning of the sixteenth century to the end of the seventeenth, the lexical range of nominalizations increased in tandem with the overall frequency of use. Judging by the fact that prominent authors appear to have led the way in broadening the range of nominalization types, it seems likely that imitation of popular and influential authors played an ever-increasing role in making nominalization a popular discursive strategy. If this was the case, the increase observed is not only the result of a change in the underlying scientific thought-style, but also of a changing writing style as a discourse communal phenomenon. Comparative synchronic studies with non-scientific genres of writing will be necessary to discover the extent to which this conjecture is accurate.

6. Conclusion

The implications of the findings presented in this study are two-fold. Firstly, evidence from the EMEMT corpus supports the qualitative claims of Halliday (1988) and Banks (2005b), showing that nominalization increased throughout the period under investigation. In our data, the turn of the seventeenth century saw the highest rate of increase. Secondly, however, we also note that although the average frequencies of nominalization show a steady increase, the linguistic practice itself was already well established from the beginning of the sixteenth century, and was utilized in select texts at a level equal to that at the end of the timeline. This observation is in line with Banks' explanation for the origin of nominalization in the scientific register, that is, translation from Latin. The early texts exhibiting a high frequency of nominalization are invariably learned translations. The high proportion of nominalizations derived from Latin verb bases also supports this argument.

Bearing in mind that the highest frequencies of nominalization in the corpus do not reach the level observed in the Newton extract – held up by Halliday and Banks as an example of the emerging nominal style – it may be that the full expressive potential of nominalizing was not fully realized in medical writing until later. The third part of the *Corpus of Early English Medical Writing*, currently under compilation, will no doubt shed light on this issue.

Notes

1 While this study was carried out, Jukka Tyrkkö was employed by the Research Unit for Variation, Contacts and Change in English (VARIENG), funded by the Academy of Finland at the University of Helsinki. The authors are members of the Scientific Thought-styles project at VARIENG, and co-compilers of the *Early Modern English Medical Texts* corpus (forthcoming). The paper has benefited from comments received at ICAME 29 from Douglas Biber, Andreas Jucker, Lilo Moessner, and Sebastian Hoffmann.

2 In this respect, nominalizations are not to be confused with 'shell nouns', a related but nonetheless separate concept. See Schmid (2000: 366).

3 Nevalainen (1999: 353) notes that nominalization resembles nominal inflection in the sense that both suffixes and inflectional morphemes alter the meaning of the base form in a relatively predictable way. However, some semantic shift also frequently takes place and thus such derivation can arguably be considered a word formation process (Nevalainen 2006: 59-63).

4 Several researchers define nominalizations simply as nouns derived from verbs and adjectives. See e.g. Reeves (2005: 39, 130) and Baker (2006: 153).

5 It may be noted that the form of some suffixes is given differently in different lists. For example, where Atkinson gives *-tion*, Nevalainen (1999) lists *-ation*. According to Bauer (2001: 181), "it seems probable that all [suffixes ending in *-ion*] should be treated as allomorphs of the same morpheme".

6 Similarly, Macleod et al. (2000: 45) use a list of suffixes in identifying nominalization for the NOMLEX lexicon of nominalizations.

7 A semantic metaphor, by contrast, involves a change of meaning without a change in word class.

8 In practical terms, Banks (2005b) notes that it is often close to impossible to distinguish between references to activities and their results.

9 The volume of tokens ending with *-ing* in the corpus is in itself prohibitive and this, together with the complexities involved with the syntactic categorization of gerunds (see Alexiadou 2001; Heyvaert 2003a), suggested it was not feasible to include nominals derived with *-ing* in the study.

10 At the time the corpus searches were conducted, the six categories comprised 1,712,307 words of running text. The number of text extracts makes it impossible to list all the titles; the list is available on request from the authors. The composition of the corpus used should be more or less identical to the final release version of EMEMT (forthcoming).

11 For a thorough discussion of the EMEMT category system, see "Manual" in the *Early Modern English Medical Texts* corpus (forthcoming). See also http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/.

12 A non-parametric ANOVA test for significance requires a minimum group size. The 'treatises' category in EMEMT is not large enough for testing until the seventeenth century.

13 The p values for the first three 50-year sections are close to 0.05, the commonly held threshold for statistical significance. However, it is prudent to keep in mind that the choice of 0.05 as the significance level (suggesting a five percent likelihood of chance occurrence) is essentially arbitrary (see e.g. Baayen 2007: 73-76 for discussion). The 6.6 and 5.2 percent values for the first and third sections, respectively, still suggest 93.4 percent and 96.8 percent likelihoods that the differences are not the result of chance.

14 The importance of different traditions or subgenres in Late Medieval medical writing has been demonstrated in numerous studies conducted with the *Middle English Medical Texts* corpus. See e.g. Pahta (2006) and Tyrkkö (2006).

15 On the differences in rhetoric between these early scientific journals, see Gross et al. (2000).

16 Judging by figure 5.4 in Atkinson (1999: 122), the corresponding frequency would appear to be approximately 12/1000 for 1675. Unfortunately no numerical data is provided to verify this.

17 The 24,000-word extract was downloaded from The Newton Project website www.newtonproject.sussex.ac.uk.

18 Halliday (1988) notes that Newton wrote *Opticks* already in the 1680's, but published it in 1704.

19 Nominalizations produced with the suffix *-ing* are also in italics, but did not count toward the frequency. See section 2.1.

20 As an isolated observation, it is interesting that nominalizations of the type *ex- + base + nominal suffix* are exceedingly rare in the sixteenth-century section of the corpus. Marchand (1969: 165) notes that "imitation" of Latin usage of the prefix *ex-* only appears in the eighteenth century.

21 Comment from the floor at ICAME 29 on May 15th, 2008 in Ascona, Switzerland.

Primary sources

anon. (1526), *Here begynneth a newe boke of medecynes intytulyd or callyd the Treasure of pore men whiche sheweth many dyuerse good medecines for dyuerse certayn dysseases as in the table of this present boke more playnly shall appere. The boke of medecines*. London: J. Rastell for Rycherd Bankes.

Bright, T. (1586), *A treatise of melancholie Containing the causes thereof, & reasons of the strange effects it worketh in our minds and bodies: with the physicke cure, and spirituall consolation for such as haue thereto adioyned an afflicted conscience*. London: Thomas Vautrollier.

Early Modern English Medical Texts (forthcoming). Compiled by I. Taavitsainen, P. Pahta, M. Mäkinen, C. Suhr, M. Ratia, T. Hiltunen, V. Marttila and J. Tyrkkö.

Gratarolo, G. (1574), *A direction for the health of magistrates and studentes Namely suche as bee in their consistent age, or neere thereunto: drawen aswell out of sundry good and commendable authours, as also vpon rea-*

son and faithfull experience otherwise certaynely grounded. London: William How for Abraham Veale.

Iorden, E. (1603), *A briefe discourse of a disease called the suffocation of the mother Written vppon occasion which hath beene of late taken thereby, to suspect possesion of an euill spirit, or some such like supernaturall power. Wherin is declared that diuers strange actions and passions of the body of man, which in the common opinion, are imputed to the diuell, haue their true naturall causes, and do accompanie this disease.* London: Iohn Windet.

Maynwaringe E. (1683), *The method and means of enjoying health, vigour, and long life adapting peculiar courses for different constitutions, ages, abilities, valetudinary states, individual proprieties, habituated customs, and passions of mind: suting preservatives and correctives to every person for attainment thereof.* London: J.M. for Dorman Newman.

References

Alexiadou, A. (2001), *Functional Structure in Nominals. Nominalization and Ergativity*. Amsterdam and Philadephia: John Benjamins.

Atkinson, D. (1999), *Scientific Discourse in Sociohistorical Context. The Philosophical Transactions of the Royal Society of London, 1675-1975*. London: Lawrence Erlbaum.

Baayen, R.H. (2007), *Analysing Linguistic Data. A Practical Introduction to Statistics.* Cambridge: Cambridge University Press.

Baker, P. (2006), *Using Corpora in Discourse Analysis*. London: Continuum.

Banks, D. (2003), 'The evolution of grammatical metaphor in scientific writing', in: L. Ravelli, A-M. Simon-Vandenbergen and M. Taverniers (eds.) *Grammatical Metaphor: Views from Systemic Functional Linguistics.* Amsterdam: John Benjamins. 127-148.

Banks, D. (2005a), 'The case of Perrin and Thomson: an example of the use of a mini-corpus', *English for Specific Purposes*, 24: 201-211.

Banks, D. (2005b), 'On the historical origins of nominalized process in scientific text', *English for Specific Purposes*, 24: 347-357.

Bauer, L. (2001), *Morphological Productivity*. Cambridge: Cambridge University Press.

Biber, D. (1988), *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Brown, P. and S.C. Levinson (1987), *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

French, R. (2003), *Medicine before Science: The Rational and Learned Doctor from the Middle Ages to the Enlightenment.* Cambridge: Cambridge University Press.

Gross, A.G., J.E. Harmon and M. Reidy (2000), 'Argument and 17th-century science: a rhetorical analysis with sociological implications', *Social Studies of Science*, 30: 371-396.

Halliday, M.A.K. (2004) [1988], 'The language of physical science', in: J.J. Webster (ed.) *The Language of Science*. London: Continuum. 140-158.

Halliday, M.A.K. and R. Hasan (1976), *Cohesion in English*. London: Longman.

Halliday, M.A.K. and J.R. Martin (1993), *Writing Science. Literacy and Discursive Power*. London: The Falmer Press.

Heyvaert, L. (2003a), *A Cognitive-Functional Approach to Nominalization in English*. Berlin: Mouton de Gruyter.

Heyvaert, L. (2003b), 'Nominalization as grammatical metaphor. On the need for a radically systemic and metafunctional approach', in: L. Ravelli, A-M. Simon-Vandenbergen and M. Taverniers (eds.) *Grammatical Metaphor: Views from Systemic Functional Linguistics*. Amsterdam: John Benjamins. 65-99.

Johns, A. (2002), 'Science and the Book', in: J. Barnard, D.F. McKenzie and M. Bell (eds.) *The Cambridge History of the Book in Britain. Volume IV: 1557-1695*. Cambridge: Cambridge University Press. 274-304.

Macleod, C., A. Meyer, R. Grishman, L. Barrett and R. Reeves (2000), 'Designing a dictionary of derived nominals', in: N. Nicolov and R. Mitkov (eds.) *Recent Advances in Natural Language Processing II*. Amsterdam: John Benjamins. 45-57.

Mäkinen, M. (2006), *Between Herbals et alia: Intertextuality in Medieval English Herbals*. Helsinki: University of Helsinki.

Marchand, H. (1969), *The Categories and Types of Present-day English Word Formation. A Synchronic-Diachronic Approach*. München: C.H. Beck'sche Verlagsbuchhandlung.

McConchie, R.W. (2002), 'Doctors and dictionaries in sixteenth-century England', in: J. Fisiak (ed.) *Studies in English Historical Linguistics and Philology: A Festschrift for Akio Oizumi*. Studies in English Medieval Language and Literature. Frankfurt am Main: Peter Lang. 267-292.

Myers, G. (1990), *Writing Biology. Texts in the Social Construction of Scientific Knowledge*. Wisconsin: The University of Wisconsin Press.

Nevalainen, T. (1999), 'Early Modern English Lexis and Semantics', in: R. Lass (ed.) *The Cambridge History of the English Language. Vol. 3, 1476-1776.* Cambridge: Cambridge University Press. 332-458.

Nevalainen, T. (2006), *Introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

Norri, J. (1992), *Names of Sicknesses in English 1400-1550: An Exploration of the Lexical Field.* Helsinki: Suomalainen Tiedeakatemia.

O'Sullivan, M.I. (1926), 'Hamlet and Dr. Timothy Bright', *PMLA* volume XL: 678-679.

Pahta, P. (2006), '*Ful Holsum and Profetable for the Bodi:* a corpus study of amplifiers in medieval English medical texts', in: M. Dossena and I. Taavit-

sainen (eds.) *Diachronic Perspectives on Domain-specific English*. Bern: Peter Lang. 207-228.

Plag, I. (2003), *Word-formation in English.* Cambridge: Cambridge University Press.

Reeves, C. (2005), *The Language of Science*. London: Routledge.

Schmid, H.-J. (2000), *English Abstract Nouns as Conceptual Shells*. Berlin and New York: Mouton de Gruyter.

Taavitsainen, I. and P. Pahta (2004), 'Vernacularisation of scientific and medical writing in its sociohistorical context', in: I. Taavitsainen and P. Pahta (eds.) *Medical and Scientific Writing in Late Medieval English.* Cambridge: Cambridge University Press. 1-23.

Tyrkkö, J. (2006), 'From *tokens* to *symptoms*: 300 years of developing discourse on medical diagnosis in English medical writing', in: M. Dossena and I. Taavitsainen (eds.) *Diachronic Perspectives on Domain-Specific English*. Bern: Peter Lang. 229-255.

Vendler, Z. (1967), *Linguistics in Philosophy*. Ithaca: Cornell University Press.

Ventola, E. (1996), 'Packing and unpacking of information in academic texts', in: E. Ventola and A. Mauranen (eds.) *Academic Writing, Intercultural and Textual Issues*. Amsterdam: John Benjamins. 153-194.

Wear, A. (1995), 'Medicine in Early Modern Europe, 1500-1700', in: L.J. Conrad, M. Neve, V. Nutton, R. Porter and A. Wear (eds.) *The Western Medical Tradition 800 BC to AD 1800*. Cambridge: Cambridge University Press. 215-361.

Wear, A. (1998), *Health and Healing in Early Modern England*. Aldershot: Ashgate.

May: The social history of an auxiliary

Arja Nurmi

University of Helsinki

Abstract

This paper explores the social history of the modal auxiliary may *between 1400-1800. The material used comes from the* Corpora of Early English Correspondence. *The development of* may *on the lexical level is tracked with regard to the social variables of gender, rank, social mobility, education and register variation. The social embedding of the rising epistemic meaning is then described with the help of a socially stratified sample of instances. The results show that the use of* may *is connected to the expression and negotiation of relative power and intimacy between correspondents. Furthermore, the rise of epistemic* may *originated with educated men, and spread gradually to other language users.*

1. Introduction

Pilot studies show that English modal auxiliaries have a social history both as lexical items and on the level of meaning (see e.g. Nurmi 2002, 2003a, 2003b).[1] This paper charts the development of *may* as a lexical item, tying the change in frequency to social variables such as gender, social status and register variation. The study then continues with a preliminary description of some of the meaning variants evident in the data, and a more detailed description of the social embedding of the epistemic meaning, which significantly increases during the time under study. Charting the general trends of development in the corpus used for more detailed studies is not only advisable but in fact necessary, since change proceeds at a different rate in different genres, and therefore results from previous studies may not give a reliable indicator of the timing of developments in the particular corpus used (Nurmi 1999: 37-38). In the case of modals this is particularly advisable since their use is tied to social context and negotiation of power and solidarity between participants in communication (see e.g. Brown and Levinson 1987).

The data used in the study come from the *Corpora of Early English Correspondence* (CEEC400; a combined version of the different parts of the corpus, totalling 5.2 million words from personal letters). The analysis of meaning is based on a socially stratified sample from the total of 16,000 instances of *may* retrieved from the corpus.

2. Previous research on *may*

Frequency-based corpus research on the history of *may* (or any modal auxiliaries, for that matter) is still thin on the ground. Kytö (1991) describes the Old and Middle English development of *can* and *may* using a pre-final version of the *Helsinki Corpus*. Her results show that by the middle of the fourteenth century *may* was only used as a modal auxiliary. Kytö (1991: 160) also attests a declining trend for *may*, but her results are problematic as a point of comparison as she treats *can* and *may* as a single variable, and presents the trend in percentages. When translated to normalised frequencies, the frequency of non-epistemic *may* in the last Middle English subperiod appears to be 1.66 per 1,000 words. Kytö (1991: 153) also finds a slightly rising trend for epistemic *may*, although it is considerably less frequent than non-epistemic. The frequency of epistemic *may* in the Middle English subcorpus, when related to the wordcount, is 0.1 per 1,000 words (slightly under five percent of the instances of *may*).

For the Early Modern English period, Kytö (1991: 204) carries out a comparison between British English (BrE) and American English (AmE). This shows a tendency for AmE to prefer deontic *may* over deontic *can* while in BrE the trend is the opposite. For BrE the frequency of *may* is highest in the text categories 'history' and 'official correspondence', and lowest in 'diary', 'private correspondence' and 'trial'. This would seem to indicate that the more oral genres favoured the new trend of declining *may* and increasing *can*, but the problem of treating *may* and *can* as a single variable continues to blur the issue. When normalised, the frequency of deontic *may* in the British private correspondence is 2.14 per 1,000 words and 2.93 in the American (Kytö 1991: 206).

On the basis of the final version of the *Helsinki Corpus*, Dury (2002) confirms this declining trend of *may* from the fourteenth to the seventeenth century. Dury also confirms that there is clear genre variation in the use of *may*, official letters having a particularly high frequency. The problem with Dury's otherwise very detailed study is the two-century gap between the periods studied. Since change proceeds at different speeds it is difficult to estimate whether a sampling period this long is sufficient to catch significant moments in the history of *may*.

Continuing the diachronic trend from 1650 to the end of the twentieth century, Biber (2004) shows that the declining trend for *may* can be seen in the ARCHER corpus as well, although there is a great deal of variation between genres.[2] BrE letters show the most clearly declining trend (3.0 to 0.9 instances per 1,000 words), while drama, for example, only starts the decreasing trend in the twentieth century and medical prose shows a steady increase until the first half of the twentieth century. These varied developments are most likely tied to the history of each genre.

Again, for twentieth-century English, Leech (2003) establishes a declining trend for *may* in the LOB and FLOB corpora across genres. His research suggests that central modals are in the process of being replaced by semi-modals. Furthermore, Facchinetti (2003) establishes social variation on the lexical level of

may in present-day spoken BrE. Her results from the *International Corpus of English - Great Britain* show that education is a significant social variable in the use of *may* while speaker's gender and age, for example, are not.

3. The social history of modal auxiliaries

There is as yet no systematic charting of the social history of English modal auxiliaries. As previous research on historical corpus-based sociolinguistics has shown, there are many new insights and viewpoints available from both stratificational and interactional sociolinguistic and socio-pragmatic studies (see e.g. Nevalainen and Raumolin-Brunberg 2003; Laitinen 2007; Nevala 2004; Nurmi 1999; Palander-Collin 1999; Nurmi, Nevala and Palander-Collin (eds.) in press).

Traditional multi-genre corpora allow for the observation of differences between genres, which can often give indications of the differences in the levels of orality and degrees of formality of linguistic variables (see e.g. Kytö 1991; Dury 2002 and Biber 2004). At the other end of the continuum is the fine-grained analysis of individual acts of communication (see e.g. Fitzmaurice 2002; Dossena forthcoming), with modality as one of the linguistic variables contributing to the negotiation of social relationships in particular situations.

It is not surprising that much of the more detailed work in historical sociolinguistics and socio-pragmatics is focussed on correspondence: the genre presents real communication between identifiable participants (for a further discussion, see Nurmi and Palander-Collin 2008). Both Dossena (forthcoming) and Fitzmaurice (2002) mine their own carefully compiled corpora for details on the microlevel of communication. My own research, while also using correspondence, is initially more concerned with the macrolevel of sociolinguistics, and with the social variables of gender and social status, as well as register variation tracked with the help of recipient types.[3] Initial results concerning *will*, *shall* and *must* (Nurmi 2002, 2003a, 2003b) indicate that this is a fruitful avenue of enquiry.

4. *Corpora of Early English Correspondence*

The material used in this study comes from the *Corpora of Early English Correspondence*. The version used here is a combination of three subcorpora: the original *Corpus of Early English Correspondence* (CEEC), the *Corpus of Early English Correspondence Extension* (CEECE) and the *Corpus of Early English Correspondence Supplement* (CEECSU).[4] The three parts of the corpus have here been united as CEEC400, covering four centuries. Table 1 shows the statistics of the separate and combined corpora. The combined number of letter writers in the corpora is 1131.

The corpora were compiled so as to be socially representative of the literate social ranks of England in 1400-1800. Surprisingly, the problems in

compiling the eighteenth-century material for CEECE were remarkably similar to those for earlier times: the availability of good editions was limited, and the letters of lower ranks of society were only sporadically edited, even where they existed in greater numbers in archives.

All three corpora have been compiled according to similar principles, with one exception. In the CEECSU, editions with modernised spelling have also been included. The letters in question have been provided with appropriate coding so that their influence on any study can be monitored. In the case of *may* (or modal auxiliaries on the whole) it seems highly unlikely that modernisation of spelling has a significant effect (if any) on the validity of results.

Table 1: *Corpus of Early English Correspondence* (CEEC400)

	CEEC	CEECE	CEECSU	CEEC400
Words	2.6 M	2.2 M	0.4 M	5.2 M
Letters	5,935	4,965	859	11,759
Collections	95	77	19	382
Timespan	1510-1681	1653-1800	1402-1663	1402-1800

Table 2: Wordcounts for different writer categories in CEEC400

	15th c.	16th c.	17th c.	18th c.	Total
Total	397,165	971,129	1,931,278	1,932,683	5,232,255
Men	331,774	894,898	1,568,660	1,346,545	4,141,877
Women	65,391	76,231	362,618	586,138	1,090,378
High rank	86,110	471,309	792,860	570,475	1,920,754
Middle rank	222,906	297,317	871,613	1,239,604	2,631,440
Low rank	88,149	202,503	266,805	122,604	680,061
FN	172,907	187,063	558,429	628,082	1,546,481
FO	27,508	144,251	326,301	170,499	668,559
FS	60,446	26,760	25,886	2,669	115,761
TC	895	7,831	269,616	521,942	800,284
T	135,409	606,108	750,521	595,913	2,087,951

The combined corpus is not equally representative of all social groups through the centuries, but it is probably representative of the body of letters produced at the time (for a discussion of representativity, see also Palander-Collin et al. forthcoming). Table 2 (above) gives some indication of this. As to gender, it is only in the late eighteenth century that the proportion of women (measured in wordcount rather than number of writers or letters) rises to half of that of men. Similarly, the lowest ranks of society are not represented nearly as consistently as the higher ranks. The three-tiered division of social ranks is based on a test study, which indicated that with regard to *may* the clearest results were obtained by grouping the social ranks into three categories, high, middle and low.[5]

Finally, recipient relationship is also unequally represented, with 'family nuclear' (FN) and 'other' (T) being well represented throughout, and 'family other' (FO) reasonably well, but 'family servant' (FS) and 'close friend' (TC) uneven, the former more in the early part of the corpus and the latter in the eighteenth century. The different recipient types are also gendered, with FS being almost completely a male category and TC including few women before the eighteenth century.

5. The social history of *may*

The overall development of *may* in personal correspondence is in line with the findings of earlier research (see figure 1 below; also table 1a in appendix). The only exception to the declining pattern is the apparent increase of *may* during the sixteenth century, which will be discussed in more detail below.

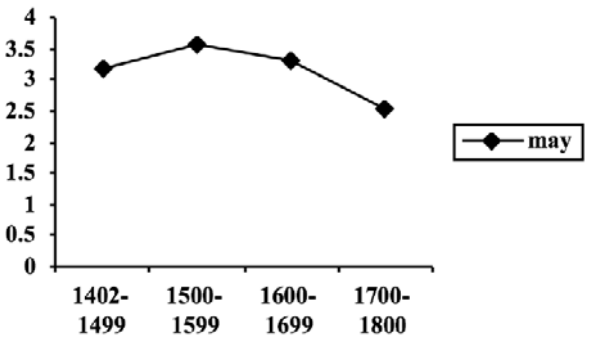

Figure 1: Development of *may* in CEEC400 (frequency per 1,000 words)

The social variables charted in this study include gender, social stratification (with a more detailed look at education and social mobility) and register variation. According to Nevalainen and Raumolin-Brunberg (2003: 185-201) gender and region are the two most robust social factors associated with linguistic change. Since the eighteenth-century part of the corpus does not allow for consistent tracking of region, this variable has been left out of the present study.

5.1 Gender

Gender is one of the social variables that shows differences in the use of almost any linguistic feature. The development of *may* exhibits a similar pattern for both genders, but the frequencies in women's letters fluctuate more. This may in part reflect the thin representation of women in the early part of the corpus, due to low literacy levels. It is also possible that the higher percentage of scribal letters in the fifteenth century obscures part of the gender difference (see also Nurmi 1999: 34-

36; Nevalainen and Raumolin-Brunberg 2003: 113-116; Meurman-Solin and Nurmi 2004: 305-307).

Figure 2 (and table 1a in the appendix) presents the results for both genders by century. It can be noted that the overall pattern for both genders is one of decline over the centuries, with the exception of the sixteenth century, where both genders show a clear increase. While the overall trend in development is highly statistically significant for both genders ($p<0.001$), the difference between men and women is only statistically significant in the eighteenth century ($p<0.001$).

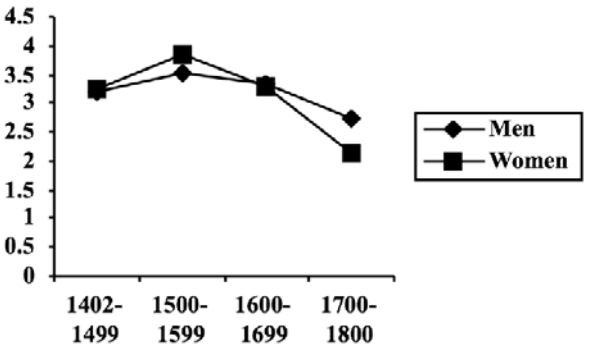

Figure 2: *May* and gender (frequencies per 1,000 words)

Women seem to lead the decreasing trend from 1700 onwards. This would indicate that the decline of *may* is a change from below (Labov 1994: 78), both in terms of the level of social awareness, and also position in social hierarchy (cf. Nevalainen and Raumolin-Brunberg 2003: 115-116). In the time period studied, women had arguably less access to prestige language simply because of unequal educational opportunities, even though in the eighteenth century the situation of the highly literate informants was beginning to change this (but see Sairio in press).

5.2 Social stratification

For the purposes of this study, the results for social stratification were initially tracked with a fairly detailed division into social ranks, following the coding in the sender database accompanying the corpora (for details, see e.g. Nurmi 1999: 39-40, Nevalainen and Raumolin-Brunberg 2003: 136-154). These were then grouped into larger categories in two alternative ways. Initially, there was a division into gentry and non-gentry informants, as in e.g. Meurman-Solin and Nurmi (2004: 307-309) and Nurmi and Palander-Collin (2008). This division turned out to be too generalised. Therefore, a three-tiered model was applied. This decision is supported by the results of Nurmi and Palander-Collin (2008: 43-44), which indicate that modals are used to express and negotiate power

imbalances. In this representation of the data, the high rank includes informants from royalty, nobility and upper gentry (knights), as well as upper clergy (bishops and archbishops). The middle rank includes the gentry, professionals (lawyers, civil servants, doctors and the like) and lower clergy. Finally, merchants and the group 'other' (including e.g. servants and farmers) form the low rank. The few informants of whom not enough is known to reliably place them in any rank category were also included in the low rank in this model.

As figure 3 below and table 2a in the appendix show, the three social ranks each follow a similar pattern in the use of *may* over the centuries. The high ranks have at all times more instances of *may* than the other two. This could be an indication that the auxiliary is used to negotiate power when writing to social inferiors (see also Nurmi and Palander-Collin (2008), which suggests that during the eighteenth century the frequency of *may* was clearly higher when writing to social inferiors than when writing to social superiors). The middle group has in most cases more frequent use of *may* than the low group, which has the lowest incidence of the auxiliary. The exception to this pattern is again the sixteenth century, where the lowest group shows higher frequencies of *may* than the middle one. All differences between social ranks and the development for each social rank are statistically significant ($p<0.05$ for rank differences in the eighteenth century and $p<0.01$ for rank differences in the sixteenth century; otherwise $p<0.001$).

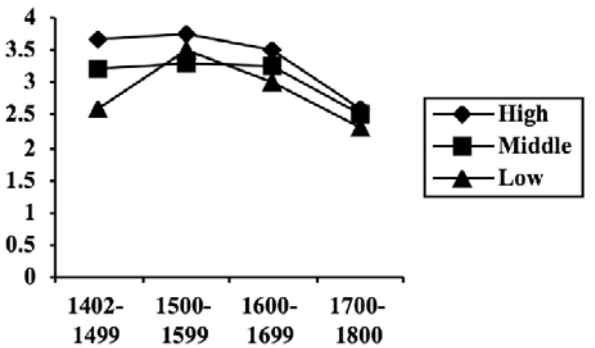

Figure 3: *May* and social stratification (frequencies per 1,000 words)

The high frequency of *may* among the low ranks in the sixteenth century seems to be mostly due to the Johnson family of wool merchants. The Johnsons were fairly affluent before their bankruptcy and had aims of gentility with the purchase of land, for example, but clearly were not members of the gentry in their station in life. The patterns observed among the family and its connections give some indication of why the frequency of *may* is higher than expected. Two frequent letter writers, Anthony Cave and John Johnson, both typically write to people who can be regarded as being in a socially inferior position to them. Most letters

by Cave are addressed to John Johnson, who, although by now a partner, was apprenticed to Cave, and continues to show deference towards his old master attributable to the difference in age and wealth, in addition to the fact of the former apprenticeship. Similarly, most of John Johnson's letters are addressed to either his younger brothers, junior partners in the family business, or to various agents of the company. Therefore, it can be argued that in the microcosm of the Johnson business circle, Cave and Johnson held the highest rank and expressed this in their frequent use of *may*. This interpretation is in accordance with the results of Nurmi and Palander-Collin (2008), where the use of most modals (including *may*) was associated with writing to social inferiors.

5.2.1 Social mobility

A subcategory of social status is social mobility, which again appears to be a significant factor in linguistic behaviour. In present-day sociolinguistics, social ambition has been measured through interviews (Chambers 1995: 95-100), but in historical studies social mobility is measured by actual social rise during an individual's life span (Nurmi 1999: 41-42; Nevalainen and Raumolin-Brunberg 2003: 150-153). In this study, social mobility is grouped into three categories: steep aspirers who rose more than two ranks in their lifetime (typically non-gentry who through education and patronage gained a title or bishop's see), mild aspirers who rose two ranks (e.g. a gentleman first knighted and then receiving a title) and those whose social status declined (e.g. younger sons of gentlemen who learned a trade).

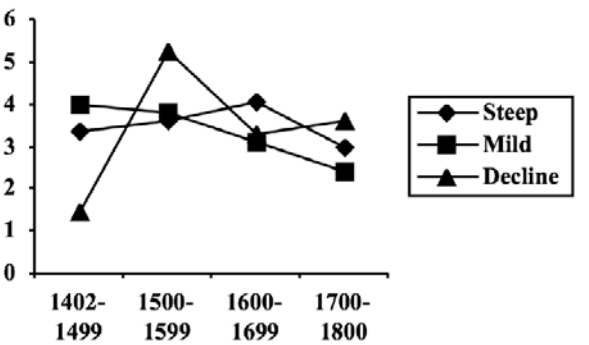

Figure 4: *May* and social mobility (frequencies per 1,000 words)

The results in figure 4 above and table 3a in the appendix present different patterns for all the three groups. Steep aspirers show a higher frequency of *may* than the corpus average, although during the sixteenth century the difference is very slight. More interestingly, steep aspirers increase their use of *may* longer than letter writers in general, with their peak in the seventeenth century. This

could be an indication of their successful social rise, although it should be noted that their use of *may* in the seventeenth century is clearly more frequent than that of the high social rank in general. It would seem that social aspirers correctly interpreted *may* as a feature of power and overused it in a way perhaps related to hypercorrection.

The group of mild aspirers follows the overall pattern more closely, except for the fifteenth century, where their frequency is higher than that of others. The declining trend after that is in accordance with the general trend, although the decline in this group is more rapid than the corpus average. This could be an indication that the mild aspirers, typically starting from a higher social position than the steep aspirers, are more in tune with the language of those born to power, and therefore follow the trends more naturally. Individual informants contributing to the high frequencies of *may* in the fifteenth century include the two Paston brothers, John II and John III. Again, it is possible that the nature of their correspondence is reflected in their use of the modal: after their father's death both in turn served as the head of the family and this may have influenced their linguistic practices.

The downwardly mobile group is unfortunately unevenly represented, and clear trends are difficult to identify. In the sixteenth century, the very high frequency of *may* is possibly an indication that the writers are attempting to maintain their old social status by linguistic means, and are in fact overcompensating when doing so. In the seventeenth century, social decline seems to have no particular influence: writers behave very much like their original rank, and are not far from the corpus average.

5.2.2 Education

Education is another social variable of interest in the context of earlier centuries, although relevant in present-day studies as well (see Facchinetti 2003, for example). Educational opportunities were limited to those with a high social position or suitable patronage. In the corpus database, high education level is coded for informants who studied at a university or the Inns of Court, but also for those privately educated to a high degree (Queen Elizabeth I being an obvious example; see Nurmi 1999: 38-39; Nevalainen and Raumolin-Brunberg 2003: 40-43). Education has not often been used as a variable to track linguistic change, but there are some indications that, particularly in the early stages of change from above, it can be a significant feature (Meurman-Solin and Nurmi 2004: 309; see also section 6.2 below). It is worth pointing out that in our sender database education was not one of the factors used to assign rank to informants, except in some cases for difficult-to-classify individuals, who were more likely to be assigned to the category 'professional' if they had a higher education. Therefore, education can be treated as an independent variable in this study.

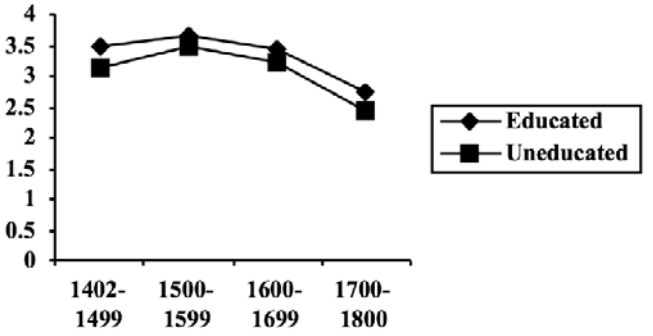

Figure 5: *May* and education (frequencies per 1,000 words)

Figure 5 above and table 4a in the appendix show the differences between the highly educated and the less educated, by century. It should be noted that even the group labelled 'uneducated' in the table usually received some level of education, formal or otherwise, since they were capable of writing their own letters. Even if the letters were written by a scribe, there was still some layer of education involved in transcribing thought to parchment or paper. The differences between the two groups are consistent in that the more highly educated group at all times uses *may* more. The difference between the two groups is statistically significant in all four centuries ($p<0.05$ during the first two, $p<0.01$ in the seventeenth century and $p<0.001$ in the eighteenth), as is the development shown for both groups ($p<0.001$). This may be an indication of the typically higher social position of the educated, but it may also be linked to the rise of the epistemic sense of *may* (section 6.2). The results shown in figure 5 definitely support the overall pattern of *may* being a feature of social power. There may also be a link to a better knowledge of social mores by the educated, regardless of their social position, something Meurman-Solin and Nurmi (2004) describe as "stylistic literacy".

5.3 Register variation

Unlike gender and social status, which are (despite social mobility) usually fairly stable variables, relationship to the interlocutor, or in the case of letters, to the recipient is constantly changing. Not only are there different recipients, requiring a different tone and style from the writer, but there are also diverse communicative situations, requiring different aspects of communication to be suited to their specific particulars. This continuum of social and linguistic behaviour is tracked in the corpus under the heading of register variation.

There are five different categories recognised in the corpus coding: 'family nuclear' (FN; parents, children, siblings and spouses), 'family other' (FO; aunts, uncles, cousins, grandparents etc), 'family servant' (FS, where the relationship is not familial, but the correspondents live in the same household and the relation-

ship produces some long-term intimacy despite the power imbalance), 'friend' (TC, close friends) and 'other' (T, all other recipients from acquaintances to total strangers). (More on register variation as used in connection with the CEEC corpus family can be found in Nurmi 1999: 43-45; Nurmi and Palander-Collin 2008; Palander-Collin 2006).

Figure 6 below and table 5a in the appendix present the results for *may* according to recipient type and century. In most cases the declining trend can be seen, although not quite as obviously as in the case of the overall results or the other social variables. The trend of development is statistically highly significant for all recipient types ($p<0.001$). The difference between the recipient types is statistically highly significant at all times ($p<0.001$), except for the seventeenth century, where it is still significant ($p<0.05$).

A comparison of family letters shows that members of the nuclear family consistently receive fewer instances of *may* addressed to them than other family members. This ties in well with the hypothesis that *may* is used to negotiate power and intimacy: when there is greater intimacy, there is less need to expressly negotiate it. Power differences between family members, on the other hand, did exist quite clearly in the time period studied, even between siblings of different ages, and these would have been suitable grounds for using *may*.

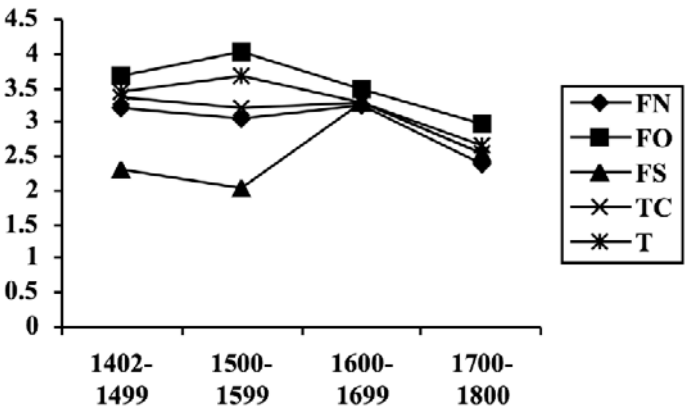

Figure 6: *May* and register variation (frequencies per 1,000 words)

A very clear power differential is at play in the family-servant relationships, on the other hand. The proportion of this category in the corpus is unfortunately uneven through the time studied, but it can clearly be seen that, on the whole, the frequency of *may* is quite low here, with the exception of the seventeenth century. This anomaly is easily explained: the code is used both for servants writing to their masters and vice versa, and while during the other centuries letters from servants predominate, in the seventeenth century the relationship is reversed, and the bulk of the correspondence in the category is from masters or mistresses to

their servants. This supports the interpretation that *may* is indeed used to negotiate power from above to below.

Letters to close friends are in many ways similar to letters addressed to nuclear family members, with a fair amount of intimacy, but also existing power differentials due to differing courses in life. Therefore, it is not surprising that the use of *may* in these situations follows the trends of family letters, staying intermediate between nuclear and other family. It seems that, in this case, as well as in the case of family members, intimacy trumps any potential power difference. While close friends would change their form of address, for example, in accordance with the social mobility of the addressees (see Nevala 2004), other linguistic practices, such as the use of modals, seem to reflect the existing intimacy.

The most heterogeneous recipient category is the group 'other', which appears to follow the overall trend observed in the corpus most closely. While the results for the other recipient types could be interpreted in terms of power or intimacy, such estimations are more difficult in this case. While, by a process of exclusion, the group 'other' does not include people with whom the writers were in a significantly intimate relationship, there is still a great deal of variation in that respect, ranging from acquaintances to total strangers. There is also a full range of power differentials from letters by kings to their subjects to begging letters by paupers to their parishes.

6. The meanings of *may*

While the previous section looked at the development of *may* on the macrolevel, as a lexical item, the levels of semantics and pragmatics should not be overlooked. My main focus is on the spread of the epistemic meaning, but I look at the other uses of *may* as well, initially grouped under the label 'deontic'. All examples and generalisations in this section refer to the socially stratified sample described in more detail in section 6.1, but I believe the description and generalisations apply to the whole data as well.

The epistemic meaning is the one easiest to single out, although there are, of course, ambiguous or unclear instances as well. Only the clearly epistemic cases were included under this label, and the ambiguous ones were included in the deontic category. Example (1) shows one of the earliest unambiguous instances of epistemic *may*.[6] The development of epistemic *may* in its social context is discussed in more detail in section 6.2.

(1) I thought and yet thinke, that it *may* be that I was shett vp againe, vpon some newe causeles suspicion, growen peraduenture vpon some secret sinister informacion. (Thomas More to Margaret Roper, 1534)

The much larger category of *may* that I have for the purposes of this study labelled 'deontic' contains some fairly clear subgroups. In example (2) Anne

Seymour refers to William Paget's ability to do something. She is flattering Paget for his wisdom, but is actually also referring to his very real position of power as the Lord Comptroller. This is one of the clear ways in which *may* is used to expressly and yet obliquely refer to power.

(2) I knowe yow *maye* do muche good in these matters beinge a wiseman. (Anne Seymour to William Paget, 1549)

Perhaps the most stereotypical use of *may* is giving or receiving permission, as in example (3). Although particularly stereotypical, it is not in fact very common, and it most frequently appears in the negative. The example here shows one of the few affirmative instances and, revealingly, yet again encodes a clear power imbalance. Christopher Lowther is writing instructions to his servant Richard Powley, and chooses to encode the master-servant relationship not only with the use of *may* in the permission sense but also with the personal pronoun *thou*. This is, according to Nevala (2004: 173), a somewhat old-fashioned use of the pronoun, since by then *thou* had become more of an indicator of intimacy than relative power, at least in family letters. The difference in recipient type does, of course, make a difference to the usage patterns as well.

(3) Thou *may* send them over by some of our neighbors. (Christopher Lowther to Richard Powley, 1639)

Deontic *may* sometimes has a meaning which can be interpreted as a fairly neutral future, and could be replaced with *will* with only very little change in meaning. In example (4) *may* seems to be a slightly more deferential choice. The deferentiality is directed at the addressee of the letter, who is in charge of seeing that a third party receives the money owed him by Elizabeth Fenwick. Again, there is a clear request, if not command, to pay the money, but the force of the demand is softened by the use of *may*, as well as the choice of other lexical items, such as "his just desire".

(4) I prey you and Mr Mikelton take care that my cousin Weston *may* be satisfied in his just disire consarning my hundered pownd bond. (Elizabeth Fenwick to Miles Stapylton, 1667)

The most semantically empty uses of *may* are formulaic. There are two formulas typical of the early material: *it may please* and *may it please*. The first, as in example (5) is used when deferentially asking for something from a person in a higher social position. The latter, as in (6), is typically a letter opening formula, also associated with deferential occasions. This formula seems to be somewhat idiosyncratic, favoured by some writers but not others in corresponding circumstances. *May it please* appears most frequently in the late sixteenth and early seventeenth centuries.

(5) *It may please* youe to sende me worde by this beyrer. (Robert Brereton to William Brereton, 1526)
(6) *Maye it please* your Ladiship I had present accesse vnto his lordship. (Philip Alpe to Katherine Paston, 1624)

As with any linguistic feature, pragmatic layers can be added to its use. So deontic *may* does not necessarily code deferentiality, but can be applied for humour, as in example (7) where Gabriel Harvey describes the joys of academic life to his father. The use of *may* is undeniably linked to a source of power (the one banishing Gawber from the table), but at the same time it makes light of the situation, in combination with the description of Gawber's dog-like behaviour.

(7) And as for gentle M. Gawber, his Mastership *may* go shake his eares elswhere, and appoint his diet at sum other table. (Gabriel Harvey to John Harvey, 1573)

Finally, there is what Dury (2002: 126) calls the "subjunctive-like" use of *may* in wishes. Example (8) shows a typical instance of this usage. These wishes seem to become more common in the eighteenth century. Very often they seem to appear where there is either a power imbalance or when the writer is being (or attempting to appear) particularly pious (see also Nurmi and Nevala forthcoming).

(8) *May* you enjoy every wish of your benevolent heart. (Ignatius Sancho to Lydia Leach, 1775)

Even this brief introduction to the many facets of the meanings of *may* clearly indicates that a more in-depth study is necessitated. Particularly, linking the uses of the auxiliary with the circumstances of each instance and the relative power and degree of intimacy between correspondents is a promising avenue of future research.

6.1 Socially stratified sample of instances

The examples of auxiliary *may* presented above and selected for semantic study were chosen on social grounds. A total of 180 instances were sampled from each century, half of them by men and the other half by women. Since the three-tiered stratification model seemed to give the best results on the lexical level, it was adopted as the basis of the stratificational sampling: 30 examples for men and the same amount for women, each selected from the low, middle and high ranks of society for each century. The third social variable chosen as a basis for sampling was register variation. The five different recipient types the corpus has been coded with, each with different degrees of intimacy and varying power dynamics, were used as sampling criteria. This meant that six instances from each stratificational sample would represent each recipient type. Table 3 gives an outline of the sampling principles.

In addition to these principles, half of the examples were, whenever possible, selected from the first half of the century in question, half from the latter half. In addition, autograph letters were given priority.

As was already seen with the results for the social embedding of the lexeme *may*, all social groups are not equally represented in the corpus. Furthermore, since there were clear differences between different user groups in the frequency of the auxiliary, this made sampling even more problematic. Table 4 presents the realised sampling of *may* from the full corpus.

Table 3: Social categories for semantic study: sampling principles

Rank	Gender	Recipient types					
		FN	FO	FS	TC	T	Total
High	Men	6	6	6	6	6	30
	Women	6	6	6	6	6	30
Middle	Men	6	6	6	6	6	30
	Women	6	6	6	6	6	30
Low	Men	6	6	6	6	6	30
	Women	6	6	6	6	6	30
Total		36	36	36	36	36	180

Table 4: Social categories for semantic study: realised sampling

Rank	Gender	Recipient types					
		FN	FO	FS	TC	T	Total
High	Men	24	18	12	18	24	96
	Women	24	24	6	12	24	90
Middle	Men	24	24	18	21	24	111
	Women	24	24	0	12	24	84
Low	Men	24	20	19	12	24	99
	Women	17	5	6	0	7	35
Total		137	115	61	75	127	515

Instead of the 720 instances ideally sampled, the corpus only yielded 515 in total. Some categories, for example 'family nuclear' and 'other' (T in the table) gave the full 24 examples in total, except for low-ranking women. In many cases, particularly for the middle ranks, I was able to follow the sampling criteria to the full, but with women in the low ranks I was reduced to taking what was available. Of the recipient types, 'family servant' and 'friend' (TC) were predictably the most difficult categories to find a full set of examples for. In fact, men in the seventeenth century were the only group where it was possible to select a full set of examples from all three ranks to all five recipient types.

Since this sampling was to some extent experimental in itself, and meant to be a test of how to approach the total of 16,000 examples in some sensible

manner, I decided to proceed with the selection. I believe that this method still provides me with a wide range of the usage of *may* in the material, and allows for further plans for more targeted studies on the meanings. Since the clearest change in progress was the spread of the epistemic sense, I decided to focus on its social embedding.

6.2 The rise of epistemic *may* in its social context

While there are some earlier isolated cases, the increase of epistemic *may* in the data begins during the course of the sixteenth century. Only the clearly epistemic instances were included in the category, the ambiguous cases were left in the much larger deontic group for now. The trends in the data agree well with earlier research, for example Dury (2002). Table 5 shows the development in the socially stratified sample. The development and differences between the two meanings are statistically highly significant ($p<0.001$). Even remembering that the sampling is not as balanced as was planned, the tendency is remarkably clear. It also corresponds well to the similar trend evident for epistemic *must* (Nurmi 2003a: 116-118). It is possible that the increasing use of epistemic *may* in the sixteenth century is also one of the contributing factors to the overall increase of the auxiliary in that century.

Table 5: Epistemic and deontic *may*

	Epistemic		Deontic		Total
	N	%	N	%	N
1402-1499	1	1	100	99	101
1500-1599	12	9	116	91	128
1600-1699	35	22	124	78	159
1700-1800	41	32	86	68	127

As to the social embedding of epistemic *may*, it shows clear symptoms of being a change from above. Men are the early adopters of the meaning, and up to and including the seventeenth century they use epistemic *may* clearly more than women. During the eighteenth century the situation changes, and women take the lead. This is parallel to the development observed for epistemic *must* (Nurmi 2003a). One of the factors explaining this development may be education, which was more available to men than women. Dury (2002: 95) notes that the earliest epistemic instances of *may* appear in religious treatises and philosophical writings. This would seem to support the role of education in spreading the epistemic usage. The early adopters in correspondence are almost without exception university-educated (and occasionally religious) men like Thomas More, Archbishop Matthew Hutton, Nathaniel Bacon and John Chamberlain. Therefore it would seem that the origins of epistemic *may* were in the academic and religious discourse practised at universities and the form was introduced into common parlance by educated men. This agrees with the results for epistemic *must* (Nurmi 2003a), which also seems to have spread from the same group.

Other social variables provide some support. The use of epistemic *may* is most frequent among the informants of high social rank and least frequent among the low ranks. This reflects the availability of higher education in the population, and supports the idea that the construction spread from the language of the highly educated men to the rest of the population. Register variation offers some indications which could also be interpreted as favouring the education hypothesis: the lowest incidence of epistemic *may* is in letters to and from family servants. This is the recipient relationship most clearly oriented to power imbalance, and least focussed on rational argumentation. The highest frequency is found in letters to the category 'other', which would seem to indicate that epistemic *may* was at its most useful when negotiating social distance and a lack of intimacy, in which case logical arguments may well have been a useful tool.

7. Conclusion

The overall development of *may* observed in the data shows a declining trend over the four centuries studied, with the exception of an increase in the sixteenth century. The social variables observed speak of stability in some cases and of change in others. The most stable difference is the most frequent use of *may* among the highest social ranks, and the corresponding least frequent use among the lowest. This seems to indicate that the auxiliary was regarded as a means of enacting and negotiating power relationships, particularly the use of power by those in possession of it towards their dependents or social inferiors. This interpretation is supported by the behaviour of social aspirers, who seem to have overcompensated in their use of *may* in their high positions. The results concerning education can be considered as giving further support, since higher education was in most cases related to either birth into a privileged social rank or a means of social advancement for the less fortunate.

Also supporting the hypothesis that *may* was a means of negotiating both power and intimacy are the results for register variation. It appears *may* is used more often as social distance between correspondents increases, whether in terms of power or emotion. Members of the nuclear family and close friends are much less often addressed with *may* than servants or strangers. Particularly the relationship between servants and masters ties in with the differing modal usage by the high and low ranks of informants since it is likely that people in high ranks will find many opportunities for writing to addressees in a position socially inferior to them, while the lowest social strata will find much more occasion for writing to their betters.

The results for gender are less conclusive, since during the first three centuries studied there is no statistically observable variation between the frequencies of *may* in letters written by men and women. A clear difference between them only appears in the eighteenth century, when women take the lead in the decline of *may*. The fact that, by the eighteenth century, women lead the

development could be seen as an indicator of this being a change from below in the Labovian sense.

In contrast, the rise of the epistemic sense of *may* from the sixteenth century onwards shows all the hallmarks of being a change from above: the form is first adopted by educated high-ranking men, and spreads later to more general use. It is likely that epistemic modality is something that would have been picked up in studies at the universities or Inns of Court, as both scholarly and legal argumentation would make use of markers of certainty and likelihood.

The social history of *may* clearly needs further study, particularly an exploration of the microlevels of interaction, which is likely to shed new light on the macrolevel patterns observed in the data. Similarly, a more detailed analysis of the meanings of *may*, and linking them to their social contexts, is a subject for future work. On the macrolevel, exploring *may* in the larger context of other modal expressions of the same meanings would also provide more insight into the patterns of variation.

Notes

1 The research reported here has been funded by the Academy of Finland, and supported by the Research Unit for Variation, Contacts and Change in English. This study is part of a larger charting of the social history of English modal auxiliaries (1400-1800).

2 Biber (2004) does not clearly indicate which version of ARCHER he has used, but mentions that the corpus size is 1.7 million words.

3 Register is not used here as a synonym of genre, but as a term for different styles and degrees of formality suited to different communicative situations.

4 The two published versions of the corpus, the *Corpus of Early English Correspondence Sampler* (CEECS) and the *Parsed Corpus of Early English Correspondence* (PCEEC) are both selections from the original CEEC. The two corpus versions are available for academic use from the Oxford Text Archive and on the ICAME corpus CD-ROM.

5 The term rank is used rather than class in this study. While the class-based society saw its beginnings in the eighteenth century, during the earlier centuries rank is a better descriptor of social status based on birth and wealth.

6 It is worth noting that Thomas More was also one of the early adopters of epistemic *must* (Nurmi 2003a: 117).

References

Biber, D. (2004), 'Modal use across registers and time', in: A. Curzan and K. Emmons (eds.) *Studies in the History of the English Language II. Unfolding Conventions*. Berlin: Mouton de Gruyter. 189-216.

Brown, P. and S.C. Levinson (1987), *Politeness. Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Chambers, J.K. (1995), *Sociolinguistic Theory*. Oxford: Blackwell.

Dossena, M. (forthcoming), 'Building trust through (self-)appraisal in 19th-century business correspondence', in: P. Pahta, M. Nevala, A. Nurmi and M. Palander-Collin (eds.) *Social Roles and Language Practices in Late Modern English*.

Dury, R. (2002), '*May*', in: M. Gotti, M. Dossena, R. Dury, R. Facchinetti and M. Lima (eds.) *Variation in Central Modals. A Repertoire of Forms and Types of Usage in Middle English and Early Modern English*. Bern: Peter Lang. 83-128.

Facchinetti, R. (2003), 'Pragmatic and sociological constraints on the functions of *may* in contemporary British English', in: R. Facchinetti, M. Krug and F. Palmer (eds.) *Modality in Contemporary English*. Topics in English Linguistics 44. Berlin: Mouton de Gruyter. 301-327.

Facchinetti, R., M. Krug and F. Palmer (eds.) (2003), *Modality in Contemporary English*. Topics in English Linguistics 44. Berlin: Mouton de Gruyter.

Fitzmaurice, S. (2002), *The Familiar Letter in Early Modern English*. Amsterdam: John Benjamins.

Kytö, M. (1991), *Variation and Diachrony, with Early American English in Focus*. Bamberger Beiträge zur Englischen Sprachwissenschaft 28. Frankfurt am Main: Peter Lang.

Labov, W. (1994), *Principals of Linguistic Change. Volume 1: Internal Factors*. Oxford: Blackwell.

Laitinen, M. (2007), *Agreement Patterns in English. Diachronic Corpus Studies on Common-number Pronouns*. Mémoires de la Société Néophilologique de Helsinki 71. Helsinki: Société Néophilologique.

Leech, G. (2003), 'Modality on the move: the English modal auxiliaries 1961-1992', in: R. Facchinetti, M. Krug and F. Palmer (eds.) *Modality in Contemporary English*. Topics in English Linguistics 44. Berlin: Mouton de Gruyter. 223-240.

Meurman-Solin, A. and A. Nurmi (2004), 'Circumstantial adverbials and stylistic literacy in the evolution of epistolary discourse', in: B.-L. Gunnarsson, L. Bergström, G. Eklund, S. Fridell, L.H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren and M. Thelander (eds.) *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe, ICLaVE 2 Uppsala University, Sweden, June 12-14, 2003*. Upp-

sala: Department of Scandinavian Languages, Uppsala University. 302-314.

Nevala, M. (2004), *Address in Early English Correspondence. Its Forms and Socio-pragmatic Functions*. Mémoires de la Société Néophilologique de Helsinki 64. Helsinki: Société Néophilologique.

Nevalainen, T. and H. Raumolin-Brunberg (2003), *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Pearson Education.

Nurmi, A. (1999), *A Social History of Periphrastic DO*. Mémoires de la Société Néophilologique de Helsinki 56. Helsinki: Société Néophilologique.

Nurmi, A. (2002), 'Does size matter? The *Corpus of Early English Correspondence* and its sampler', in: H. Raumolin-Brunberg, M. Nevala, A. Nurmi and M. Rissanen (eds.) *Variation Past and Present. VARIENG Studies on English for Terttu Nevalainen*. Mémoires de la Société Néophilologique de Helsinki 61. Helsinki: Société Néophilologique. 173-184.

Nurmi, A. (2003a), 'The role of gender in the use of MUST in Early Modern English', in: S. Granger and S. Petch-Tyson (eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. Amsterdam: Rodopi. 111-120.

Nurmi, A. (2003b), '*Youe shall see I will conclude in it*: sociolinguistic variation of WILL/WOULD and SHALL/SHOULD in the sixteenth century', in: D. Hart (ed.) *English Modality in Context. Diachronic Perspectives*. Linguistic Insights, Studies in Language and Communication 11. Bern: Peter Lang. 89-107.

Nurmi, A. and M. Nevala (forthcoming), 'The social space of an eighteenth-century governess: modality and reference in the private letters and journals of Agnes Porter', in: P. Pahta, M. Nevala, A. Nurmi and M. Palander-Collin (eds.) *Social Roles and Language Practices in Late Modern English*.

Nurmi, A., M. Nevala and M. Palander-Collin (eds.) (in press), *The Language of Daily Life in England (1400-1800)*. Amsterdam: John Benjamins.

Nurmi, A. and M. Palander-Collin (2008), 'Letters as a text type: interaction in writing', in: M. Dossena and I. Tieken-Boon van Ostade (eds.) *Studies in Late Modern English Correspondence*. Bern: Peter Lang. 21-49.

Palander-Collin, M. (1999), *Grammaticalization and Social Embedding*: *I THINK and METHINKS in Middle and Early Modern English.* Mémoires de la Société Néophilologique de Helsinki 55. Helsinki: Société Néophilologique.

Palander-Collin, M. (2006), '(Re)constructing style and language as social interaction through first- and second-person pronouns in Early Modern English letters', in: I. Taavitsainen, J. Härmä and J. Korhonen (eds.) *Dialogic Language Use*. Helsinki: Société Néophilologique. 339-362.

Palander-Collin, M., M. Nevala and A. Nurmi (in press), 'The language of daily life in the history of English: studying how macro meets micro', in: A.

Nurmi, M. Nevala and M. Palander-Collin (eds.) *The Language of Daily Life in England 1400-1800*. Amsterdam: John Benjamins. 1-23.

Sairio, A. (in press), 'Methodological and practical aspects of historical network analysis: a case study of the Bluestocking letters', in: A. Nurmi, M. Nevala and M. Palander-Collin (eds.) *The Language of Daily Life in England 1400-1800*. Amsterdam: John Benjamins. 107-135.

Appendix

Table 1a: *May* and gender (f = normalised frequencies per 1,000 words)

Time	Men		Women		Total	
	N	f	N	f	N	f
1402-1499	1056	3.18	211	3.23	1267	3.19
1500-1599	3161	3.53	294	3.86	3455	3.56
1600-1699	5204	3.32	1192	3.29	6396	3.31
1700-1800	3660	2.72	1243	2.12	4903	2.54
Total	13081		2940		16021	

Table 2a: *May* and social rank (f = normalised frequencies per 1,000 words)

Time	High		Middle		Low	
	N	f	N	f	N	f
1402-1499	318	3.69	719	3.23	230	2.61
1500-1599	1761	3.74	984	3.31	710	3.51
1600-1699	2768	3.49	2820	3.24	808	3.03
1700-1800	1491	2.61	3130	2.52	282	2.30

Table 3a: *May* and social mobility (f = normalised frequencies per 1,000 words)

Time	Steep aspirers		Mild aspirers		Social decline		Corpus average	
	N	f	N	f	N	f	N	f
1402-1499	16	3.34	287	4.00	9	1.46	1267	3.19
1500-1599	371	3.59	599	3.77	117	5.24	3455	3.56
1600-1699	770	4.02	1253	3.07	671	3.29	6396	3.31
1700-1800	14	2.95	825	2.42	18	3.60	4903	2.54

Table 4a: *May* and education (f = normalised frequencies per 1,000 words)

Time	Educated		Uneducated	
	N	f	N	f
1402-1499	229	3.47	1038	3.13
1500-1599	1837	3.64	1618	3.47
1600-1699	3135	3.44	3261	3.20
1700-1800	1796	2.74	3107	2.43

Table 5a: *May* and register variation (f = normalised frequencies per 1,000 words)

Time		Family nuclear	Family other	Family servant	Friend	Other
1402-	N	557	101	139	3	467
1499	f	3.22	3.67	2.30	3.35	3.45
1500-	N	574	581	54	25	2221
1599	f	3.07	4.03	2.02	3.19	3.66
1600-	N	1811	1140	100	881	2464
1699	f	3.24	3.49	3.86	3.27	3.28
1700-	N	1487	510	1	1328	1577
1800	f	2.37	2.99	0.37	2.54	2.65

Go to V: Literal meaning and metaphorical extensions

Sara Gesuato

University of Padua

Abstract

Motion verbs can literally encode goal-oriented motion (come to the office*;* come to the meeting) *or metaphorically express a change of state* (come to a decision)*, including when they are followed by non-finite complements* (come to say goodbye*;* come to conclude).

Analysis of about 1,400 corpus concordances reveals that go to V *mainly encodes its motional meaning, in association with human subjects (88 percent;* Jeremy goes to answer the door*) and, occasionally, four related figurative ones: 'other-determined transfer or use (of a resource)' in association with inanimate subjects (three percent;* and everything went to pay debts*); 'contribution to an outcome' when collocating with verbs denoting involuntary processes (two percent;* there are too many factors that go to make up a great marathon runner*); 'succeeding (in demonstrating)', if co-occurring with pronominal subjects identifying inanimate, abstract entities, and with the verbs* prove *or* show *(six percent;* and it just goes to prove anybody can play*); and 'going on or proceeding (with a course of action)' in association with human subjects performing deliberate acts not involving physical motion (one percent;* we then have to go to develop the job into […] a strategy*).*

Acceptability judgements expressed by twelve native speakers on 20 sample sentences show that the motional meaning of go *is also applicable to the construction variant* be going to V *if the scenarios represented express short-term goals.*

The motional meaning of go *thus appears to be synchronically relevant to its verbal complements in specific lexico-syntactic environments, allowing metaphorical extensions outside the domain of tense.*

1. Introduction

Expressions literally encoding spatial notions can be used metaphorically to represent temporal-aspectual concepts. This is because space is a fairly concrete domain of experience, and therefore can serve as the basis for the interpretation and expression of the more abstract domain of time. This involves the mapping of concepts and formulas from the former, source domain onto the latter, target domain, that is, the setting up of a systematic analogy, or metaphorical correspondence, between the two domains (Lakoff and Johnson 1980). To give but a

simple example, to contextualize an event with respect to some reference time, it is possible, and often necessary, to resort to spatial prepositions which effectively *locate* the event *in* time, and which represent the specific reference time as if it were a spatial entity, that is, as a one-dimensional point (e.g. *at noon*), a two-dimensional surface (e.g. *on Monday*) or a three-dimensional container (e.g. *in winter*).

A temporal concept which is often understood and represented in spatial terms is the "course of event" (McIntyre 2001: 150). This is interpreted as a path, and relevant encoding formulas involve the association of a motion verb with an adverbial expressing a metaphorical destination. That is, the motion verb is syntactically used in its typical pattern, i.e. followed by an adverbial, but the meaning of the construction in which it occurs is metaphorically extended to express the achievement of a goal; e.g.

(1) his lack of success leads me to the conclusion that the only people being given offers of articles are those with personal introductions (BoE: N0000000381)[1]
(2) I'll never get to the stage where that doesn't give me goose bumps (BoE: N0000000571)
(3) Add the stock powder, curry powder and cream, with the sugar, salt and pepper and bring to the boil (BoE: N5000950118)
(4) He came to the decision that football would take precedence (BoE: N5000950406)

A similar metaphorical extension may take place when motion verbs are followed by non-finite verbal complements. In such cases, they metaphorically express transition from one state to another, and function as lexical aspectualizers (Brinton 1988) encoding such notions as (causative) resultativity, inception or completion; e.g.

(5) his inspection did not lead him to believe there was a 'problem' (BoE: N5000950307; causative resultativity)
(6) I got to know the two top divers (BoE: N5000950301; inception)
(7) She also got to keep the 24 or more dresses he made for each film (BoE: N0000000806; resultativity)
(8) You can't bring yourself to believe I've forgiven you (BoE: B9000001423; causative resultativity)
(9) The longer you lived, the better you came to understand the quirks of human nature (BoE: B0000000298; prospective culmination)

Over time, the metaphorical use of a motion verb may become its default, most commonly instantiated meaning, so that its original, literal meaning becomes obscured in the process.[2] The verb, that is, undergoes a process of grammaticalization: it stops conveying a spatial meaning, it develops a temporal meaning,[3] and this new meaning becomes relevant to one of the organizing principles of the

language system (e.g. a realization of a category in the structure of the language). This is what appears to have happened to the verb *go*.

The verb *go* literally expresses wilful, physical, translational motion away from a source location (Goddard 1997). However, like other motion verbs, it can be used metaphorically to denote a process of transition from one situation to another (Talmy 1975: 234); this occurs when it is accompanied by adverbials identifying outcomes, results or target states; e.g.

(10) But promise me you won't let your nerves go to pieces (BoE: B9000001423)
(11) They've let the ground go to ruin in recent years (BoE: N9119980511)
(12) you had nightmares when you went to sleep (BoE: S0000000783)

In addition, when occurring in a progressive form (i.e. *be going*) and followed by a *to*-infinitival complement, it encodes the notion of 'projected realization of an event with respect to a reference time', possibly coloured by such semantic nuances as intentionality/premeditation, predictability/inevitability, imminence or current relevance; e.g.

(13) it was going to cost them more (BoE: S90000001513)

Indeed, the temporal usage of *be going to V* (i.e. futurity) has become so strongly grammaticalized that the motional meaning of *go* is hardly perceived in that syntactic environment.[4]

Finally, *go* retains its motional meaning when followed by *V-ing*, at least with verbs denoting routinized errands, social events or institutionalized activities; e.g.

(14) So what do you do when he goes shopping? (BoE: S9000001301)
(15) I once went camping in May (BoE: S9000001248)

Extensive research has been carried out on the usage of infinitival clauses and catenative constructions in English (e.g. Eastlack 1967; Fang 1995; Hoekstra 1988; Mair 1990; Stevens 1972; Tortora 1988; Whelpton 2001, 2002; Egan 2008) and of the *be going to V* construction in particular (e.g. Binnick 1971; Bybee and Pagliuca 1987; Haegeman 1989; Hopper and Traugott 2003; Danchev et al. 1965; Danchev and Kytö 1994; Mair 1997; Nicolle 1997, 1998a, 1998b; Facchinetti 1998; Poplack and Tagliamonte 1999; Krug 2000; Brisard 2001; Szmrecsanyi 2003; Berglund and Williams 2007). However, the focus of the research has always been on the encoding of tense and aspect. As far as I know, no study has been carried out on the literal usage of *go* when followed by *to*-infinitives. This paper has two aims: to explore in what lexico-syntactic environments *go to V* constructions instantiate the meaning of goal-directed motion (on the basis of corpus data), and to check whether this meaning is also applicable to progressive

be going to V constructions (on the basis of native speakers' acceptability judgements).

2. Collection and analysis of corpus data

To examine the motional usage of *go* with non-finite complements, I searched the BoE for occurrences of the forms *go*, *goes*, *gone* and *went* followed by *to V*, *to be V-ed* and *to be V-ing* (i.e. active non-progressive, passive non-progressive and active progressive *to*-infinitives). That is, I looked for instances of *go to*-infinitives that are not known to encode temporal meanings by default.

I excluded from the query output concordances exemplifying ambiguous or irrelevant constructions, such as: sequences of *go to* followed by a word of unclear part of speech (e.g. *go to sleep/work*); sequences of *go to* preceded by place adverbials interpretable as encoding locations or destinations (e.g. "I asked him where I should go to eat" (BoE: N0000000680): 'I asked him what place we should go to in order to eat' versus 'I asked him in what place we should eat'); and sequences of *go to* occurring as part of larger constructions (e.g. "the lengths to which he went to offer them a way back" (BoE: N2000960213)).

Table 1: Distribution of *go*, *goes*, *gone*, *went* across three non-finite complements

Go forms	to V	to be V-ed	to be V-ing	Global
go	238 (17%)	4 (57%)	0 (0%)	242 (18%)
goes	143 (10%)	0 (0%)	0 (0%)	143 (10%)
gone	105 (8%)	1 (14%)	0 (0%)	106 (8%)
went	881 (64%)	2 (29%)	0 (0%)	883 (64%)
Total	1,367 (100%)	7 (100%)	0 (0%)	1,374 (100%)

Table 2: Distribution of the concordances in spoken and written texts

Construction	Spoken	Written	Global
go to V	96 (40%)	142 (60%)	238
go to be V-ed	2 (50%)	2 (50%)	4
goes to V	24 (17%)	119 (83%)	143
goes to be V-ed	0 (0%)	0 (0%)	0
gone to V	33 (31%)	72 (69%)	105
gone to be V-ed	0 (0%)	1 (100%)	1
went to V	302 (34%)	579 (66%)	881
went to be V-ed	1 (50%)	1 (50%)	2
Total	458 (33%)	916 (66%)	1,374

Table 1 above shows the frequency of the forms of *go* and their distribution over the three non-finite complements considered. The most frequent verb-form is

went (accounting for 64 percent of the data); the second most frequent is *go*, relevant to 17 percent of the occurrences; *goes* and *gone* together account for the remaining 18 percent of the concordances. All the four forms strongly correlate with the occurrence of *to V* complements (i.e. active non-progressive *to*-infinitives).

On average, most of the occurrences (66 percent) are found in written texts. The preference for the written medium applies to all the verb-forms considered, and is especially high in the case of *goes* (83 percent; see table 2 above).

A variety of tenses are exemplified in the corpus (see table 3), but only the simple past stands out as prominent (74 percent of the time). The simple present, the second most frequent tense, accounts for eleven percent of the data, while the present perfect and past perfect are instantiated in seven percent of the concordances. The use of auxiliaries is infrequent: the two most frequent ones are *would* and *will*, with 19 and eleven occurrences, respectively. Examples of the variant realization of *go to V* / *be V-ed* follow:

(16) Dr Kilduff went to book a room at a chateau hotel (BoE: N2000951128)
(17) I did not go to change anything in the house (BoE: B0000001245)
(18) when you go to fetch your supermarket trolley (BoE: S0000000093)
(19) The money goes to provide fresh, clean water (BoE: N0000000381)
(20) I don't even go to buy bread or milk any more (BoE: N0000000048)
(21) I will say I don't regret having gone to bury myself in the provinces (BoE: S0000000832)
(22) He's gone to get some ice (BoE: B0000000140)
(23) he would simply assume she'd gone to help Jill (BoE: B0000000906)
(24) Mr Beron said his party would now go to consult with its electorate (BoE: S1000901102)
(25) players will go to ease the massive wages bill (BoE: N6000920513)

Table 3: Distribution of the concordances across tenses

Tense	Frequency	Tense	Frequency
Simple present	157 (11%)	will future	17 (1%)
Simple past	1,012 (74%)	imperative	7 (1%)
Present perfect	42 (3%)	bare infinitive	18 (1%)
Past perfect	50 (4%)	to-infinitive	37 (3%)
Present conditional	15 (1%)	past to-infinitive	3 (0.4%)
Past conditional	4 (0.6%)	other	12 (1%)
All tenses			1,374 (100%)

Table 4 below shows that, of the three non-finite complements considered, *to V*, with 1,367 occurrences, accounts for over 99 percent of the data, while *to be V-ed* is relevant to less than one percent of the occurrences, and to *be V-ing* is not attested at all.

Table 4: Frequency of three non-finite complements of *go*

Non finite complements	Occurrences	Percentage
to V	1,367	99.5%
to be V-ed	7	0.5%
to be V-ing	0	0.0%
Total	1,374	100%

The non-finite complements instantiate a variety of lexemes. On average, each lexeme is exemplified four times, the value of token/type ratio being the highest in the case of *went to V* (see table 5).

Table 5: Tokens and types of lexemes in non-finite complements

Construction	Tokens	Types	Ratio
go to V	88	238	2.7
goes to V	38	143	3.7
gone to V	51	105	2.0
went to V	191	881	4.6
GO to be V-ed	6	7	1.2
Total	374	1,374	3.7

The most frequent lexeme exemplified in the corpus is *see*, both in the sense of 'visually perceive' and in the sense of 'visit' (391 occurrences), followed by *live* (79), s*how* (65), *get* (50), *visit* (48) and *meet* (29). Table 6 below shows the number of occurrences of the most frequent lexemes. Only four have 50 or more instantiations, and only eleven have 20 or more. The verbs express a few basic notions: 'meeting/visiting' (*meet*, *see*, *visit*), 'inhabiting a place' (*live*, *stay*), 'visually (intentionally) perceiving' (*look at*, *see*, *watch*), 'running an errand' (*answer*, *buy*, *collect*, *get*, *look for*, *pick up*, *take*, *tell*), 'demonstrating that something is the case' (*prove*, *show*). A few others (i.e. *help*, *make*, *investigate*, *make up*, *go*, *pay*) form a semantically heterogeneous group; e.g.

(26) we don't go to see Clare so much now (BoE: E00000002146)
(27) If I find a book hard going I often go to see the film (BoE: N9119980426)
(28) Les was forced to go to live in an old camper van (BoE: N9119980608)
(29) when you go to get a book it's never there (BoE: S9000000523)
(30) His remarks only go to show that if you say something enough times it must be true (BoE: N6000950306)
(31) And then I went to visit my brother (BoE: S9000001326)
(32) So in the end I went to meet him (BoE: S9000000765)
(33) whenever I've gone to watch a game (BoE: N9119980428)
(34) They went to look for some treasure (BoE: S9000000338)

(35) we've produced all the things that go to make civilisation, science, art and all that (BoE: N0000000740)
(36) As I went to go around it, it seemed to move (BoE: N6000940427)

Table 6: Number of occurrences of the most frequent lexemes

Lexeme	Frequency	Lexeme	Frequency
see	391	investigate	16
live	79	help	16
show	65	collect	15
get	50	prove	13
visit	48	take	13
meet	29	answer	12
pick up	27	go	11
watch	24	pay	11
buy	23	make up	8
look (at)	23	tell	7
stay	20	look for	9
make	17		

The construction appears to encode a few related meanings (see table 7 below). Most of the concordances can be interpreted as expressing goal-oriented motion (88 percent of the time). This meaning can be glossed as 'deliberately moving away towards a physical destination in order to V'. In this meaning, the construction typically patterns with human subjects – and infrequently with non-human animate subjects – and represents events encoding both long- and short-term goals.[5] The relevant verbs denote both durative and instantaneous processes (e.g. *see*, *live*, *show*, *get*, *visit*, *meet*, *pick up*, *watch*, *buy*, *look (at)*, *stay*); e.g.

(37) Jeremy goes to answer the door (BoE: B9000001423)
(38) by then he had gone to live in Cornwall (BoE: B0000000888)
(39) But police who went to check failed to spot her (BoE: N9119980520)
(40) These birds soon went to nest after I supplied them with wicker nest baskets (BoE: N0000000770)

A related meaning that the construction can encode is that of goal-directed transfer or use of a resource (three percent of the time). In this meaning, *go* retains its original motional sense, and colours it with the nuance of other-determined arrival at a (metaphorical) destination (cf. *channel*). The construction can therefore be taken to mean 'serve, be used to V / be V-ed'. When conveying this meaning, the construction patterns with inanimate subjects, often identifying financial resources (73 percent), and correlates with verbs encoding the notions of 'favouring' or 'helping financially' (i.e. *aid*, *answer*, *benefit*, *build*, *compensate*, *cover*, *develop*, *ease*, *feed*, *finance*, *fund*, *help*, *improve*, *make*, *meet*, *pay*, *pay off*, *supplement*, *supply*, *support*); e.g.

(41) Rhino horn, tiger bones and bear gall bladders all go to supply the market for oriental medicines (BoE: E0000001695)
(42) In low-tax countries investment goes to build businesses and employment (BoE: N2000951104)
(43) Bits have gone to build other bits, leaving little reminders (BoE: N0000000337)
(44) My father had just died and everything went to pay debts (BoE: B0000000140)

Table 7: Distribution of the concordances across five meanings

Construction	Meaning: 'move so as to'	Meaning: 'be transferred and/or used'	Meaning: 'contrib-ute to'	Meaning: 'succeed in'	Meaning: 'proceed to'
go to V	197	15	17	1	8
goes to V	47	15	4	76	1
gone to V	102	2	1	0	0
went to V	858	12	4	1	6
GO to be V-ed	6	0	1	0	0
Total	1,210 (88%)	44 (3%)	27 (2%)	78 (6%)	15 (1%)

A very similar meaning instantiated by the construction is that of unconscious involvement in a process (i.e. a state or situation) leading to a result or bringing about a consequence. In this meaning (relevant to two percent of the data), *go* can act as a near-synonym of *help* or *end up*, and be glossed as 'contributing to an outcome' or 'reaching a target result-location', respectively. In such cases, the construction often patterns with inanimate subjects (70 percent of the time), and it is associated with verbs denoting unconscious, involuntary processes, which convey the notions of 'constituting' or 'contributing' (i.e. *build*, *emphasize*, *make*, *make up*, *save*, *support*, *form*, *pay*, *produce*, *be replaced*, *suffocate*); e.g.

(45) the killing of the young boys and girls who are to go to make up the future strength of the Indian people is the saddest part of the whole affair (BoE: B0000091417)
(46) and there are too many factors that go to make up a great marathon runner (BoE: S1000900927)
(47) They were motivated by these elements of "deterrence", "retribution", and "moral reform" which went to make up Victorian penal-welfare strategies (BoE: B0000000854)
(48) the frantic style which can only go to suffocate the glorious talents of many top players (BoE: N6000950817)

(49) and a few of the innermost cells go to form what's called the primitive streak (BoE: S0000000285)
(50) what we go to pay for when we get back (BoE: S9000001266)
(51) Oil extracted from the plant and its seed went to produce a floor-covering called linoleum (BoE: N600941007)

The verb *go* followed by a *to*-infinitive acquires a specific meaning in association with verbs conveying the notion of attesting, namely *show* and *prove*. The meaning can be glossed as 'succeeding in (providing evidence)' or 'serving to (demonstrate that something is the case)'. In such cases (six percent of the time), *go* is a near-synonym of *succeed* or *manage*; it collocates with subjects denoting inanimate entities – most of which are pronominal (i.e. 91 percent; e.g. *it*, *this*, *that*, *(all of) which*) – and half the time it combines with the adverbs *only* (ten percent) or *just* (51 percent); e.g.

(52) And it just goes to prove anybody can play (BoE: S9000000904)
(53) His remarks only go to show that if you say something enough times it must be true (BoE: N6000950306)
(54) Which only goes to show – love stands the test of time (BoE: N6000940228)
(55) COBRA's bellicosity didn't last, but it went to show that Paris was herself again, the heart, the hub (BoE: N0000000474)

The final meaning I found instantiated in the concordances can be glossed as 'going ahead with a course of action'. This meaning accounts for only one percent of the data. It involves the use of *go* as a synonym of *proceed*, and its collocation with human subjects denoting people and with verbs denoting deliberate actions that do not require physical transfer to be carried out (i.e. *consult*, *create*, *develop*, *double*, *fill*, *flirt*, *form*, *grapple*, *grip*, *head*, *help*, *reach*, *rest*, *say*);[6] e.g.

(56) you need to be quite a high level of programming to go to create software packages (BoE: S9000000524)
(57) we then have to go to develop the job into <ZF1> a <ZF1> a <ZF0> a strategy (BoE: S9000000383)
(58) He goes to say "You are unwilling to come to me that you may have life" (BoE: S9000001022)
(59) before he went to grapple with the dark tides of his destiny (BoE: B9000000492)

Given their different lexical and semantic associations, the meanings described above can be easily identified in context. The second and third meanings are quite similar, and indeed can be instantiated with some of the same lexemes (e.g. *build*, *make*, *pay*). However, they can be kept distinct because they are compatible with different paraphrases. The second meaning is relevant to the exploitation of a

resource that is used for a purpose, the infinitive encoding deliberate, other-determined goals (possible paraphrases: 'be used in order to / serve to'). With the third meaning, the events represented appear as unavoidable endpoints or consequences of given processes; in such cases, the infinitive has a culminative-resultative nuance, and its aspectual role is emphasized when the verbs in the infinitive encode unintentional experiences (possible paraphrase: 'in the end (X happens/-ed)').

Corpus data, therefore, shows that non-progressive forms of *go* followed by *to V* / *be V-ed* mainly encode goal-oriented motion, and also that this basic motional meaning can be metaphorically extended to express related concepts of 'transfer, use, contribution'. What remains to be seen is whether the meaning of goal-oriented motion is also compatible with progressive *going* followed by *to*-infinitives. I address this question with the help of data elicited from native speakers.

3. Elicitation and analysis of informant data

Be going to V / *be V-ed* activates by default the temporal meaning of 'futurity' due to its well-established grammaticalization (see section 1). To investigate its possible compatibility with non-temporal interpretations such as those identified in section 2 for non-progressive forms of *go* + *to V* / *be V-ed*, I decided to elicit acceptability judgements from native speakers on corpus-based instances of *be going to V* that could activate a motional interpretation of *going*.

Given the high prominence of the meaning of 'goal-oriented motion' (88 percent) and the high frequency of active, non-progressive *to*-infinitives (99.5 percent) among the concordances examined, I chose to submit to native speakers sample sentences instantiating only this meaning and containing only instances of *to V* infinitives.

To prepare sample sentences exemplifying the construction *going to V* that could plausibly express the notion of goal-oriented motion, I took into consideration the verbs associated with this meaning which are the most frequently instantiated in the corpus. The verbs with 15 or more occurrences are: *see* (390), *live* (79), *show* (65), *get* (50), *visit* (48), *meet* (30), *pick up* (27), *watch* (24), *buy* (22), *stay* (20), *investigate* (16), *play* (15), *look at* (15) and *collect* (15). From this set I excluded four verbs: *live* and *stay*, which express acts tied to long-term goals; *show*, which is almost exclusively attested in the sense of 'succeeding in (demonstrating)', and which is virtually always restricted to the occurrence of the verb-form *goes* (97 percent); and, finally, *investigate* because it expresses a meaning relevant to very specific domains of experience (i.e. scientific research, and law and order). In the end, I had a list of ten verbs that I could use to prepare sample sentences with: *buy*, *collect*, *get*, *look at*, *meet*, *pick up*, *play*, *see*, *visit* and *watch*.

Next, for each of the ten verbs above I picked ten concordances at random from corpus data, and then selected two in particular, which exemplified motion

directed to the achievement of short-term goals (e.g. *go to get some grocery food*) rather than long-term plans (e.g. *go to get a master's degree*). Where necessary, I also slightly modified the concordances so that they would all instantiate single-clause sentences with first-person singular subjects and present progressive predicates.

I e-mailed the 20 sample sentences to twelve native speakers of English (one Canadian, three US American and eight British), and asked them to write, for each sentence, whether it could be an appropriate reply to the question *Where are you going?* by choosing an answer from a closed set of options: *Yes*, *No* and *Undecided*. In my directions to the consultants, I specified that *Where are you going?* is not the same question as *What are you going to do?* and I invited them to add comments, if they chose to.

The list of sample sentences and a summary of the native speakers' replies are given in table 8 below. All the sample sentences received positive acceptability ratings from most consultants (i.e. 82 percent of the time). In particular, items 7 and 8 were found acceptable by all informants, while items 1-6 and 13 were rated the same way by all informants but one. The items rated as the least acceptable were 10 and 20 (with 60 percent of positive judgements) and 12 and 19 (with 67 percent of positive judgements).

The acceptability ratings might have been even higher, had it not been for the fact that a few informants objected to the phrasing of some sentences (some of them explicitly stated so in their comments).[7] Occasional comments were given on the content of the test items, revealing that the difficulty of envisaging an appropriate context of production and reception for a given sample sentence affected its acceptability judgement.[8]

Explicit comments on the issue directly addressed by this study came from three people. One informant expressed the opinion that for the items where there is no indication of the place where the various actions are to be performed it is still possible to infer (i.e. recover from the co-text) that the speaker has some specific destination in mind. He also added that for all test items except 18 the question *Where are you going?* can be re-interpreted as 'Why are you going?'. Another found the items not containing reference to a destination or location unacceptable, but added that these could probably become acceptable if one could presume the interactants knew the places where the various actions would take place. The third one wrote that for most of the items *Where are you going?* could be easily understood to mean 'What are you going to do?'. These comments suggest that the ability to think of a spatio-temporal context in which the event being represented is likely to take place may be largely responsible for the activation of a motional meaning in instances of *going to V*.

Table 8: Acceptability ratings of the motional meaning of *going* in 20 sentences

Item	Yes	No	Undecided	Total
I am going to …				
1. … buy some champagne	11	1	0	12
2. … buy a dress	11	1	0	12
3. … collect the car	11	1	0	12
4. ... collect my jeans and T-shirt	11	1	0	12
5. … get the tape recorder	11	1	0	12
6. … get help	11	1	0	12
7. … look at the garden	12	0	0	12
8. … look at what the rest of the place looks like	12	0	0	12
9. … meet my old buddies at a tavern	9	2	1	12
10. … meet a senior official	7	2	3	12
11. … pick up my children	10	1	1	12
12. … pick up a few things	8	3	1	12
13. … play ping-pong	11	1	0	12
14. … play with Christine	10	2	0	12
15. … see a play	10	1	1	12
16. … see a doctor	9	2	1	12
17. … visit my parents	9	2	1	12
18. … visit the spot where Roland Ratzenberger died on Saturday	9	2	1	12
19. … watch my boyfriend play	8	3	1	12
20. … watch rugby	7	4	1	12
Total	197 (82%)	31 (13%)	12 (5%)	240 (100%)

4. Conclusion

Motion verbs are typically used with place adverbials to express physical movement from a source location to a target destination (e.g. *come to the office*; *come to the meeting*). But their default use can also be expanded both semantically and syntactically: on the one hand, their literal meaning can be metaphorically extended to express transition from one situation to another (i.e. non-physical movement or change of state; e.g. *come to a decision*); on the other, both their literal and their metaphorical meaning can be instantiated when they are accompanied by clausal complements (e.g. *come to say goodbye*; *come to conclude*). The expression of metaphorical movement with a clausal complement determines the encoding of temporal and/or aspectual meanings (cf. Gesuato forthcoming on *come to V*).

The verb *go* encodes literal motion when accompanied by a phrasal adverbial encoding a destination (e.g. *go to the library*). It also encodes the temporal meaning of futurity, if it occurs in a progressive form and is followed by a clausal non-finite complement (e.g. *be going to read*). This paper explored whether it could also encode its literal motional meaning when followed by a *to*-infinitive.

Corpus data has shown that non-progressive *go* very often encodes the motional meaning of voluntary movement away from a source location when followed by (typically active) *to*-infinitives, which act as clauses of purpose encoding goals (i.e. metaphorical destinations). The encoding of the meaning of goal-directed motion, which is relevant to 88 percent of the data examined, strongly correlates with the presence of subjects having human referents and with the representation of events denoting deliberate actions, and evokes reference to target destinations which, however, remain unexpressed.

In addition, corpus data has shown that non-progressive *go to V* / *to be V-ed* can express related notions (i.e. metaphorical extensions of the concept of 'physical motion'), which can be grouped under the general heading of 'contributing to an outcome'. These meanings – which are, however, infrequently instantiated – include:

- 'go on', 'set (out) to' or 'proceed', which is activated in association with human subjects seen as performing voluntary acts, but not engaged in physical motion (one percent);
- 'help', 'contribute to', 'constitute', 'end up', which is relevant to both animate and inanimate subjects, and followed by verbs mostly denoting involuntary, maybe unconscious, and possibly causal, processes (three percent);
- 'be transferred' and 'be used', which correlates with inanimate subjects typically referring to money and with predicates representing 'causal' actions expressing the idea of 'financially determining or favouring a result' (three percent);
- and 'succeed in (demonstrating)', which is typically instantiated with pronominal subjects denoting inanimate, abstract entities and only combines with the verbs *show* and *prove* (six percent).

Acceptability judgements elicited from native speakers on corpus-based data (i.e. slightly modified concordance lines) have revealed that the motional meaning of *go* can also be very frequently activated for sequences of progressive *going* followed by *to V*. The suggestions for improvement offered by a few informants on the wording of the sample sentences, and their occasional voicing of doubts about the content of the sentences themselves, suggest that a revision of the questionnaire – which might include, among other things, a clear specification of the contextual scenarios in which the various sentences might be uttered – could produce even higher acceptability ratings for the various items.

Overall, the study suggests that the motional meaning of *go* is still synchronically relevant to its clausal complements in specific lexico-semantic environments, and that it can also be compatible with *be going to V*, an otherwise highly grammaticalized construction specialized in the encoding of futurity. That

is, specific contexts can be though of – i.e. those relevant to the performance of short-term goals – that make the activation of the motional meaning of *go* quite plausible also in that syntactic environment. The findings also show that *go* allows metaphorical extensions of its literal meaning, which, however, remain outside the domain of tense; that is, the notion of 'physical motion' can be expanded to include that of 'transfer, use and/or contribution'.

One interesting expansion of this study would consist in the exploration of the possible non-accessibility of the literal meaning of physical motion in sequences of *go to V* which represent involuntary events (e.g. *I'm going to breathe polluted air*; *I'm going to grow big and strong*) or acts whose performance is not dependent on the reaching of a task-specific physical destination (e.g. A. *Where are you going?* B. *I'm going to my room to cry* / **I'm going to cry* / ?*I'm going to give him a piece of my mind*).

Notes

1 Here and elsewhere, BoE stands for *Collins Cobuild Bank of English Online*, a 57,000,000-word corpus of Present-day English.

2 See Brinton (1988) on the historical development of Present-day English aspectualizers from original motion verbs.

3 *Temporal* can mean relevant to either tense or aspect; see Klein (1994).

4 The motional meaning can be noticed in specific contexts, such as when the construction is associated with a human subject and the encoding of a deliberate act, when it is relevant to an interactional scenario involving or implying physical motion (e.g. a previous speaker's turn like *Where are you going?*), and possibly, when it is interrupted by the insertion of an expression encoding a target destination (e.g. *I'm going [to the theatre] to get tickets for tomorrow's show).*

5 Subjects identifying animate non-human entities occur only three times.

6 One subject denotes an institution, namely *organization*.

7 For instance, some consultants disliked words like *buddies* and *tavern* (item 9 in table 8); others found *my children* funny, where they would have expected *the children* (item 11); two found indefinite noun phrases problematic, namely *a tavern* in item 9 and *a doctor* in item 16; and one suggested replacing *to see a play* with *to the theatre* in item 15. Other consultants, while positively rating the sample sentences, recommended rewording them: one added the word *basketball* at the end of item 19, and one suggested inserting *badminton* instead; one suggested using *look for a dress* instead of *buy a dress* in item 2; one would have preferred *get* to *collect* in item 4; one modified *I am* into *I'm*; another suggested deleting *I am going to* in item 4 and have the sentence start with the *to*-infinitive; one

objected to the indefiniteness of the place mentioned in item 9. In short, these occasional objections were made to aspects of the wording of the original corpus concordances the sample sentences were adapted from.

8 One consultant wrote that the acceptability of items 10 and 13-20 depends on where you supposedly meet the person who asks you *Where are you going?* (for item 10, it could be, for example, in front of the office building). Another consultant wrote that he could not think of a circumstance when somebody would utter item 10. The same person found item 17 odd because of its supposed reference to a long trip (as signalled by *visit*), which made it an unusual answer to the question *Where are you going (right now?)*. Still another wrote that items 12, 19 and 20 were too generic, while item 18 was too specific.

References

Berglund, Y. and C. Williams (2007), 'The semantic properties of *going to*: distribution patterns in four subcorpora of the British National Corpus', in: R. Facchinetti (ed.) *Corpus Linguistics 25 Years On*. Amsterdam and New York: Rodopi. 109-122.

Binnick, R. (1971), '*Will* and *Be Going To*', in: Chicago Linguistic Society (ed.) *Papers from the Seventh Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago University Press. 40-51.

Brinton, L.J. (1988), *The Development of English Aspectual Systems*. Cambridge: Cambridge University Press.

Brisard, F. (2001), '*Be going to*: an exercise in grounding', *Journal of Linguistics*, 37: 251-285.

Bybee, J.L. and W. Pagliuca (1987), 'The evolution of future meaning', in: A. Giacalone Ramat, O. Carruba and G. Bernini (eds.) *Papers from the 7th International Conference on Historical Linguistics*. Amsterdam and Philadelphia: John Benjamins. 109-122.

Danchev, A. and M. Kytö (1994), 'The construction *be going to + infinitive* in Early Modern English', in: D. Kastovsky (ed.) *Studies in Early Modern English*. Berlin and New York: Mouton de Gruyter. 59-77.

Danchev, A., A. Pavlova, M. Nalchadjan and O. Zlatareva (1965), 'The construction *going to + inf.* in Modern English', *Zeitschrift für Anglistik und Amerikanistik*, 13: 375-386.

Eastlack, C.L. (1967), 'Catenative verbs in Portuguese and English: a contrastive study', *Estudios Lingüísticos*, 2(1-2): 43-56.

Egan, T. (2008), *Non-finite Complementation. A Usage-based Study of Infinitive and –ing Clauses in English*. Amsterdam and New York: Rodopi.

Facchinetti, R. (1998), 'Expressions of futurity in British Caribbean Creole', *ICAME Journal*, 22: 7-22.

Fang, A.C. (1995), 'Distribution of infinitives in contemporary British English: a study based on the British ICE Corpus', *Literary and Linguistic Computing*, 10(4): 247-257.

Gesuato, S. (forthcoming), 'Encoding of goal-directed motion in the *COME* + infinitive construction', in: A. Renouf and A. Kehoe (eds.) *Corpus Linguisitcs Reassessed* (provisional title). Amsterdam and New York: Rodopi.

Goddard, C. (1997), 'The semantics of coming and going', *Pragmatics*, 7(2): 147-162.

Haegeman, L. (1989), '*Be going to* and *will*: a pragmatic account', *Journal of Linguistics*, 25: 291-317.

Hoekstra, T. (1988), 'Small clause results', *Lingua*, 74: 101-139.

Hopper, P.J. and E. Traugott (2003), *Grammaticalization*. 2nd ed. Cambridge: Cambridge University Press.

Klein, W. (1994), *Time in Language*. London and New York: Routledge.

Krug, M. (2000), *Emerging English Modals: A Corpus-based Study of Grammaticalization.* Berlin and New York: Mouton de Gruyter.

Lakoff, G. and M. Johnson (1980), *Metaphors We Live By*. Chicago: The University of Chicago Press.

Mair, C. (1990), *Infinitival Complement Clauses in English. A Study of Syntax in Discourse*. Cambridge: Cambridge University Press.

Mair, C. (1997), 'The spread of the GOING-TO-future in written English: a corpus-based investigation into language change in progress', in: R. Hickey and S. Puppel (eds.) *Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak.* Berlin: Mouton de Gruyter. 1537-1543.

McIntyre, A. (2001), 'Argument blockages induced by verb particles in English and German: event modification and secondary predication', in: N. Dehé and A. Wannen (eds.) *Structural Aspects of Semantically Complex Verbs*. Berlin, Frankfurt and New York: Peter Lang. 131-164.

Nicolle, S. (1997), 'A relevance-theory account of *be going to*', *Journal of Linguistics*, 33: 355-377.

Nicolle, S. (1998a), '*Be going to* and *will*: a monosemous account', *English Language and Linguistics*, 2: 223-243.

Nicolle, S. (1998b), 'A relevance theory perspective on grammaticalization', *Cognitive Linguistics*, 9: 1-35.

Poplack, S. and S. Tagliamonte (1999), 'The grammaticization of *going to* in (African American) English', *Language Variation and Change*, 11(3): 315-342.

Stevens, W.J. (1972), 'The catenative auxiliaries in English', *Language Sciences*, 23: 21-25.

Szmrecsanyi, B. (2003), 'BE GOING TO versus WILL/SHALL. Does syntax matter?', *Journal of English Linguistics*, 31(4): 295-323.

Talmy, L. (1975), 'Semantics and syntax of motion', in: J.P. Kimball (ed.) *Syntax and Semantics*, volume 4. London: Academic Press. 181-238.

Tortora, C.M. (1988), 'Verbs of inherently directed motion are compatible with resultative phrases', *Linguistic Inquiry*, 29(2): 338-345.

Whelpton, M. (2001), 'Elucidation of a telic infinitive', *Journal of Linguistics*, 37(2): 313-337.
Whelpton, M. (2002), 'Locality and control with infinitives of result', *Natural Language Semantics*, 10: 167-210.

Passive constructions in Fiji English: A corpus-based study

Carolin Biewer

University of Zurich

Abstract

Inner circle varieties of English seem to show variation in the usage of the get-*passive, i.e. in constructions such as* He got hurt. *It is claimed that the* get-*passive is more frequently used in American English than in British English, while Australian English and New Zealand English hold an intermediate position (Sussex 1982: 90; Hundt 1998: 78; Hundt et al. 2008: 327f). As an alternative strategy the* be-*passive can be used* (He was hurt). *It is equally interesting to check usage, form and frequency of* be-*passives to see whether some regional variation can be found.*

Fiji English is a variety of English as a second language spoken in Fiji both by Fijians and Indo-Fijians. Previous studies on concord patterns, perfect constructions and the mandative subjunctive (Biewer 2008a, 2008b, forthcoming) suggest that Fiji English has been developing under the influence of second language acquisition, Fijian, angloversals and the exonormative influence of inner circle varieties of English, in particular New Zealand English. This paper will focus on get-*passives and* be-*passives in Fiji English to gain some insight into the differences and similarities in the usage of the passive in Fiji English, British English and New Zealand English. Data will be taken from a preliminary version of ICE-Fiji. A review of the progress of the compilation of ICE-Fiji will be given and its (current) suitability for such a study will be considered. The results will be discussed as a further step towards a corpus-based description of the grammar of Fiji English.*

1. Introduction

Fiji English is a variety of English used as a second language on the Fiji islands. It differs from Standard British English in many ways. First, there is the influence of the local substrate languages, which can best be seen in the vast number of lexical borrowings from Fijian, such as *sulu* 'wrap-around' and *bilibili* 'bamboo raft', and from Hindi, such as *babu*, 'mate' (Geraghty et al. 2006: 47, 54, 38). As for accent, Fijians, for instance, speak Fiji English with a trilled or flapped /r/ in accordance with the way they pronounce the /r/ in Fijian (Tent and Mugler 2004b: 756). There are also many interesting morpho-syntactic differences, for instance the omission of past tense or past participle suffixes as in *Pool close!* or *He decide to come*. Other examples are pronominal copy after an overt subject

noun phrase (*My mother (s)he told me...*), *one of* + sg noun (*She is one of the expert in pottery*) and the use of non-count nouns as count nouns (*They ate lots of seafoods*).[1]

Recent studies on morpho-syntactic features of Fiji English (Biewer 2008a, 2008b, forthcoming) suggest that these features have been developing under, among other factors, the influence of second language acquisition and substrate influence. On the other hand, it is the exonormative influence of inner circle varieties of English, in particular New Zealand English, which seems to form this new variety of English in the South Pacific.

This paper will focus on *get*-passives and *be*-passives in Fiji English to gain some insights into the differences and similarities of the usage of the passive between Fiji English, British English and New Zealand English. Do social hierarchies and the politeness system in Fiji call for more passive constructions in Fiji English than in inner circle varieties of English? Are more complex forms avoided because of the learning process? What kind of interference from Fijian – in which passive constructions are grammatically quite different from English (Churchward 1973: 19f) – can be detected? And is it possible to see New Zealand English replace British English as a new model for an emerging national standard of Fiji English in terms of passive constructions? Data for this study was taken from three different components of the *International Corpus of English* (ICE): a preliminary version of the Fijian component (ICE-Fiji) was compared with data from the British component (ICE-GB) and the New Zealand component (ICE-NZ).

I will divide this paper into five parts. First, I will briefly consider the usage of English in Fiji and possible influences on the development of the morpho-syntax of Fiji English (section 2). In section 3, I will discuss passive constructions in inner circle varieties and Fijian before I move on to section 4, in which information will be given on the current status of ICE-Fiji. Next (section 5), I will discuss three different case studies on *get*-passives and *be*-passives in Fiji English before I interpret the results (section 6).

2. Fiji English as an outer circle variety of English

2.1 English in Fiji

As Fiji was a British colony from 1879 to 1970, English is widely used in Fiji (Fischer 2002: 151f), particularly as a lingua franca between the two largest ethnic groups, the Fijians and the Indo-Fijians, who comprise about 95 percent of the whole population (Tent 2001: 210; Tent and Mugler: 751). English is the official language of parliament, is widely used in literature and the media and is often the sole medium of instruction in school (Mangubhai and Mugler 2003: 371f; Tent 2001: 211f; Tent and Mugler 2004b: 751). English is also usually used at the workplace, if not in conversation then at least in written correspondence. Different levels of English – acrolectal, mesolectal, basilectal – can be distin-

guished according to social class, where people grew up and to which school they went. Ethnic differences between Fijian and Indo-Fijian usage of English are claimed to be minimal on the acrolectal level (Tent and Mugler 2004b: 752). Most people speak English as a second language. There are a few part-Europeans who speak it as a first language; their English, however, closely resembles the English Fijians speak as these part-Europeans strongly identify with their Fijian upbringing (Tent and Mugler 2004b: 753). English is omnipresent in the cities, where most people live. The vast majority of the population can communicate in English and does so whenever necessary.

2.2 Major influences on the morpho-syntax of Fiji English

Grammatical features like *one of* plus singular noun (*one of the expert*) may result from substrate influence as in Fijian nouns are not marked for number with bound morphemes; numerals can be used instead (Dixon 1988: 27; Churchward 1973: 14f). However, as English in Fiji is mainly spoken as a second language, constructions such as *one of the expert* can also be interpreted as a simplification of rules at the beginning of the learning process and an avoidance of redundancies (Sand 2005: 181). Learners of English as a second language (ESL) have difficulties with bound morphemes, even if the system in their first language is similar to the system in English (Winford 2003: 218). Learners are also inclined to leave bound morphemes out as what they want to communicate can easily be grasped without them (Sand 2005: 124). With *one of* it has already been made clear that the speaker is talking about one person from a group of people, which makes the bound morpheme redundant to some extent.[2]

Differences from standard British English and American English may further arise from a reorientation of Fiji English towards New Zealand English as a new model for an emerging national standard. Today English is seen as a pluricentric language (Leitner 1992; Clyne 1992) which has several national standards developing alongside each other, which all finally develop their own endocentric norm rather than continuously following American English and British English as an exocentric norm. New Zealand English has reached that stage of "endonormative stabilisation" (Schneider 2003: 249, 269) and is now capable of influencing other newly emerging varieties of English as their new exocentric norm. New Zealand is not only geographically close to Fiji but is also a major trading partner of the Fiji Islands.[3] It has greatly influenced the educational system (Mangubhai 1984: 187) and also tries to be of influence in politics. The long-running New Zealand TV soap *Shortland Street* on Fiji television shows New Zealand's continuing presence in the media. Apart from that, the number of immigrants from Fiji living in New Zealand has risen steadily since 2001 (*Statistics New Zealand* 2002: 5). All these factors make an influence of New Zealand English on Fiji English plausible.

Recent studies on the present perfect (Hundt and Biewer 2007; Biewer 2008a), concord patterns (Biewer 2008b) and the mandative subjunctive (Biewer forthcoming) suggest that substrate influence, second language acquisition and exonormative models of competing national standards are the three major factors

that play a role in the development of Fiji English grammar. In the following, we will look at yet another feature of Fiji English grammar, namely passive constructions, to see what kind of differences and similarities we find when comparing Fiji English with British English and New Zealand English and how these findings can best be interpreted. First, the usage of the passive in inner circle varieties and in Fijian will be discussed.

3. Passive constructions in inner circle varieties of English and Fijian

3.1 The *be*-passive in inner circle varieties of English

Passive constructions are interesting from a pragmatic point of view as they are more than just a transformed active. In a short passive, in which the agent is left out – and less than 20 percent of passive clauses have a *by*-agent (Swan 1995: 410) – a human agent may not be revealed because the speaker does not know his/her identity, does not wish to reveal his/her identity or believes this information to be irrelevant (Slewierska 1984: 237). This is a "deliberate vagueness" (Slewierska 1984: 237), which "leaves the hearer with the impression that the deleted agent carried no real meaning when in fact it did" (Granger 1983: 41). This is done, in particular, in scientific writing as part of a tradition to show objective detachment (Biber et al. 1999: 477). The reason for the usage of the short passive in the media may be the "desire to save space" or to disguise a lack of knowledge of the particulars of the described event (Biber et al. 1999: 477).

As for long passives, in which an agent is mentioned, the different word order in comparison to the active sentence puts the patient in front and the agent in second position, so that the agent receives particular emphasis. To put the agent in the centre and the patient on the periphery is also a wilful decision (Granger 1983: 39, referring to Bolinger 1977). This, for instance, is an interesting strategy for the media to "maximize" the focus on "what is novel" (Biber et al. 1999: 477). These two functions of the passive are called "impersonalisation" and "topicalization" by Slewierska (1984: 218, 237).

Because of its function of impersonalisation the passive is claimed to be more common in writing than in speech and seen as a marker of formal discourse (Nash 1980: 140-142; Granger 1983: 45). It is particularly linked with text types in which impersonal style is favoured (Granger 1983: 44; Leech and Svartvik 1975: 258-259) such as academic writing, in contrast to imaginative prose, which has a particularly low frequency of passives. Leech and Smith (2006) have shown that the number of passives in American English and British English is decreasing and that this on-going grammatical change is more extreme in American English than British English. This regional variation may be caused by a different policy adopted in American style manuals or automated grammar checkers (Leech and Smith 2006: 194).

There is also another motivation to investigate the *be*-passive. When looking at mandative subjunctives in Fiji English, Biewer (forthcoming) found that

the number of passive subjunctives is particularly high in Fiji English in comparison to American English, British English and New Zealand English.[4] These occur in exactly those contexts in which the Fijian speakers wished to depersonalise their request for a certain person to do something in order not to be seen as impolite. Churchward states in his *New Fijian Grammar* that politeness strategies in the Fijian language require that orders are more often stated in the passive than in the active voice so that they have a less demanding and more polite tone (Churchward 1973: 22). Traditionally, in Fijian society hierarchies have to be preserved and the rank of one's interlocutor has to be respected in conversations. In general, learners tend to apply the politeness strategies that hold for their mother tongue to the language they are learning (Odlin 1994: 51f). A certain substrate or socio-cultural influence on passive constructions in Fiji English seems therefore possible. In that respect, it will be interesting to check whether there is a higher frequency of *be*-passives in Fiji English than in British English or New Zealand English.

3.2 The *get*-passive in inner circle varieties of English

Another verb that can be used as a passive auxiliary is *get*. The *get*-passive – in contrast to the *be*-passive – is one of the "faster-spreading recent grammatical innovations in English" (Mair 2006: 111) that is more common in speech than in writing; it was linked for the first time with a more colloquial style by Poutsma (1926-1929: 100; also see Granger 1983: 192). From a pragmatic point of view the *get*-passive shows implications that the *be*-passive lacks as it refers to an emotional involvement of the speaker (Granger 1983: 196), which often reflects an unfavourable attitude towards the action (*How did that window get opened?*, example from Quirk et al. 1985: 161). It generally refers to the unexpected nature of the described event (Granger 1983: 196) and it puts more emphasis on the (usually unfavourable) condition of the patient rather than on the agent (*He got taught a lesson / He got caught by the police*). In his study of *Twentieth-century English*, Mair found that these semantic and stylistic constraints for the *get*-passive are now lessening (Mair 2006: 113).[5] This shows that the *get*-passive can be used in contexts that originally required the *be*-passive.[6]

Inner circle varieties of English seem to show variation in the usage of the *get*-passive. It is claimed that the *get*-passive is more frequently used in American English than in British English, with the level of its usage in Australian English and New Zealand English falling between that of American English and British English (Sussex 1982: 90; Hundt 1998: 78; Hundt et al. 2008: 327f). Hundt (forthcoming) also found *get*-passives to be more frequent in Asian Englishes than in British English. The reason, however, is far from clear as this difference might be more of a stylistic difference than a regional one. The increase may be due to a colloquialisation of the written norm (Hundt 1998: 79; Mair 2006: 185ff.), in which American English is in the lead.[7] It is not clear whether American English influences the other varieties in this or whether it is a parallel and independent development and not really a regional divergence (cf. Hundt et al. 2008: 328). It will be interesting to see whether Fiji English positions itself

between British English and New Zealand English with the usage of the *get*-passive.

3.3 The passive in Fijian

Like many other languages, Fijian marks the verb in the passive with a special form, while the object of the active sentence becomes the subject of the passive sentence (cf. examples in Churchward 1973: 19f). Fijian has three different strategies to construct a passive (with slightly different functions). One strategy is to change the last vowel of the transitive suffix of a verb to *-i* as in (1):

(1) Eratou gunu-va na yaqona.
3PL drink-trs. DET kava
'They drink the kava.'

E gunu-vi na yaqona.
T/A drink-passive DET kava
'The kava is drunk.'[8]

Another possibility is to reduce the verb to its root as in (2) in which the verb *caka-va* 'to make or do' becomes *caka* 'to be made or be done':

(2) Eratou sa caka-va totolo.
3PL T/A do-trs. quickly
'They have done it quickly.'

Sa caka totolo.
T/A done quickly
'It has been done quickly.'[9]

For a limited number of verbs it is also possible to use certain prefixes to create a passive (Churchward 1973: 20). Note that for all these strategies there is no change of word order and that it is not possible to indicate the agent in a passive construction (Milner 1990: 97). If you want to say *The kava was drunk by them* you will have to put this in two sentences: *The kava was drunk. They drank the kava.*[10] All passives are constructed with the help of bound morphemes, there are no free morphemes added to the sentence when being transformed into a passive sentence (Milner 1990: 96, 97; Churchward 1973: 20, 21). It will be interesting to see whether Fiji English has fewer agentive or long passives due to the fact that there are no agentive passives in Fijian.

This gives us three different hypotheses for passive constructions in Fiji English:

1. There is a higher usage of *be*-passives in Fiji English than in the inner circle varieties;
2. There is a lower usage of agentive passives in Fiji English than in the inner circle varieties;

3. Fiji English positions itself between British English and New Zealand English in the usage of the *get*-passive.

As Fiji English is a variety of English as a second language, one can also expect to find a smaller number of complex forms, such as a combination of aspect and voice, in ICE-Fiji.

4. Methodology: a corpus-based approach

The ICE project aims at creating comparable corpora of different varieties of English around the world of 1,000,000 words each: 600,000 words of speech and 400,000 words of writing. Table 1 shows the intended number of words per text type in the ICE corpora:

Table 1: Intended number of words per text type in the ICE corpora

Categories ICE written	Approx. no. of words
student essays	40,000
correspondence	60,000
academic writing	80,000
non-academic writing	80,000
instructional writing	40,000
press	60,000
creative writing	40,000
TOTAL	400,000

ICE-Fiji is being compiled by Hundt and Biewer. The spoken component is not yet ready to be used for research, which is why, unfortunately, a comparison between speech and writing in the different ICE corpora is not yet possible. But there is enough data for most of the different text types in the written component to allow for a study on the passive across registers in comparison to ICE-GB and ICE-NZ.[11] Table 2 shows the number of words per text type in ICE-Fiji as of April 2008.

The sections 'student essays', 'press' – which contains news reports and editorial writing – are finished and the section 'creative writing' (fiction) is short of three texts but otherwise finished. There is no data for 'correspondence' at the present moment as Fiji has an oral tradition and people do not often write letters to each other. Emails will be used at some stage to fill this category. For this study, however, this category has to be excluded for all varieties. Academic writing contains more words than required as the texts have not yet been cut down to 2,000 words, which is the normal policy.

There is still a shortage of data in non-academic writing and instructional writing but, nevertheless, the data add up to around 290,000 words. For two of the three following studies all these text categories will be used except correspon-

dence; for one study only the press section will be used. For the exact number of words per text category in all three corpora see table 1a in the appendix.

Table 2: Number of words per text type in ICE-Fiji

Categories ICE written	ICE-Fiji
student essays	40,448
correspondence	0
academic writing	91,160
non-academic writing	24,890
instructional writing	23,310
press	79,105
creative writing	33,750
TOTAL	292,663

5. Three case studies

What Bauer once stated as a paradox in corpus linguistics, namely that a corpus is always either too big or too small (1994: 50f), certainly applies if one looks at passive constructions: there are either too many or too few passives to look through, which is why three different methods were used here. First, the number of *get*-passives for the different text types of the written component of the ICE corpora was checked to get enough hits. Second, the number of *be*-passives in the press section was checked, to get an impression of usage of *be* in form and function with a manageable amount of hits to search through. Finally, to be able to look at *be*-passives in the different text types of the written component of the ICE corpora, six frequently used verbs were chosen in accordance with Mair (2006), namely the verbs *tell*, *take*, *kill*, *put*, *pay* and *make*.

5.1 Defining the variable

For the study of the *get*-passive WordSmith Tools was used to retrieve all *get* + past participle. Instances of reflexive meaning such as *get washed*, *get dressed*, resulting copula such as *getting old*, *get confused* and idiomatic expressions such as *get rid of*, *get fed up with* were eliminated.

Quirk et al. describe the *be*-passive as constructions of *be* + past participle which have a clear correspondence with an active verb phrase or clause (1985: 167). These so-called 'central passives' have to be distinguished from semi-passives which show adjectival properties such as *to be interested*, *to be surprised*, *pleased*, *fed up* and pseudo-passives which denote a resultant state (Quirk et al. 1985: 167ff.). These are extremely difficult to distinguish from central passives. In cases such as *The theatre was closed* it is not clear whether the speaker refers to the state of the theatre having been closed or the action of

closing. Potential pseudo-passives were checked carefully, they were included if a dynamic reading was plausible, and excluded if a statal reading seemed more likely or if it could not be decided which reading was more likely.

WordSmith Tools again was used to retrieve all different forms of *to be* including contracted forms. Whether a participle was following and whether the *be + ed* participle was a central passive were manually checked. As it is possible in New Englishes that the participle is left unmarked (cf. e.g. Sand 2005: 93ff.; Biewer 2008a), it is not possible to check for *-ed* automatically, for instance with the context function in WordSmith. The results follow.

5.2 Case study (1): *get*-passives in the three ICE corpora

Table 3 shows the number of *get*-passives in the chosen text types of the three ICE corpora. The Mossé index gives the number of occurrences per 10,000 words in relation to corpus size.

Table 3: *Get*-passives in the three ICE corpora

get-passives	ICE-Fiji	ICE-GB	ICE-NZ
academic	1	1	3
non-academic	2	2	7
student essays	2	0	1
instructional	4	0	2
press	4	0	5
creative writing	2	5	6
TOTAL	15	8	24
Mossé	0.513	0.213	0.639

The overall number is not high enough to say much about variation across text types. But in general it can be seen that in ICE-NZ the less formal text types such as non-academic writing and creative writing and even the press section seem to have more *get*-passives than the more formal text types of academic writing, student essays (in which students imitate academic writing) and instructional writing. In ICE-GB, at least the difference between informative and imaginative prose seems obvious. The findings are also in line with previous studies that show there are more *get*-passives in the New Zealand English sample than in the British English sample (e.g. Hundt 1998: 78). The overall number of *get*-passives in ICE-Fiji positions Fiji English between British English and New Zealand English as expected. But it is far from clear whether this is a regional difference or not. Nor is it clear that this result emerges from a possible influence of New Zealand English on Fiji English. However, there is a clear difference between the inner circle varieties and Fiji English in the usage of *get* according to text type: in ICE-Fiji the *get*-passives are less common in imaginative prose and more common in instructional writing. The numbers are too small to draw a tenable conclusion but a number of possible explanations come to mind: the text type ‘creative writing’

is differently defined in Fijian culture than in the West and that may be similar with instructional writing. A Fijian novel, for instance, can include personal correspondence and non-fictitious diary entries of the author. It may also have something to do with the sampling as an attempt was made to fulfil the request stated in the ICE manual not to include texts in fiction which are too colloquial (Greenbaum 1991: 4), and other teams might have been less strict. On the other hand, instructional writing in ICE-Fiji, for which a particular regulation in the ICE manual did not exist, contains instructions that aim at a younger readership. A possible higher usage of *get*-passives in that category too might result from a different sampling strategy. The differences in the usage of the *get*-passive in the three varieties may rather be stylistic than a matter of regional divergence. The differences may mirror socio-cultural differences between the Fijian society and the West. But one has to be cautious about how representative and how comparable the ICE corpora really are or can be. So, yes, Fiji English does position itself between British English and New Zealand English in the usage of the *get*-passive but not because of an exonormative influence of the inner circle varieties.

5.3 Case study (2): *be*-passives in the press section of the three ICE corpora

Table 4 shows the absolute and relative frequencies of *be*-passives in the press section of the three ICE corpora. The figures are remarkably similar. They demonstrate that Fiji English does not have more *be*-passives than the inner circle varieties.

Table 4: *Be*-passives in the press section of the ICE corpora

be-passives in press	absolute figures	Mossé
ICE-Fiji	767	0.97
ICE-GB	622	0.95
ICE-NZ	724	1.05

Next, the different forms of *be*-passives were checked, i.e. the number of *be*-passives with a *by*-agent, the number of progressive passives and perfect passives and the number of central modals with a *be*-passive.

Figure 1 gives the percentages in relation to all *be*-passives in the respective press sections. Whereas the number of central modals with a *be*-passive is similar in all three press sections, the number of agentive or long passives in ICE-Fiji press is considerably lower than in the other two corpus sections. This could have something to do with the fact that Fijian does not have an agentive passive. It could also be a matter of avoiding more complex forms. The considerably lower number of passive progressives and perfect passives in ICE-Fiji press compared to the sub-corpora of the inner circle varieties also seems to hint at an avoidance of the more complex combination of aspect and voice, which could be a learner influence. Moreover, the higher number of *by*-agents in British English and New Zealand English actually resulted from a higher number of present

perfect and past perfect forms with *by*-agent. And that definitely shows that complexity of form is a crucial factor here.[12] So, yes, Fiji English has fewer agentive passives, as expected, but it is first and foremost a matter of second language acquisition. However, as the learner in acquiring a second language leans on the structure of his or her mother-tongue substrate influence cannot be ruled out entirely.

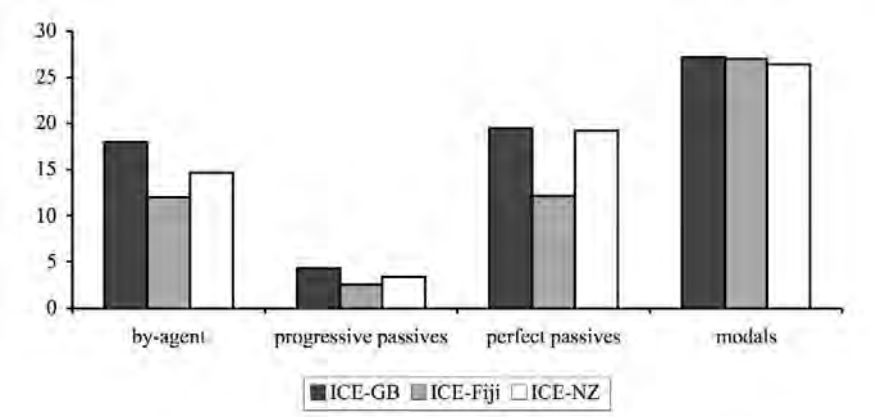

Figure 1: Percentage of *by*-agents, progressive passives, perfect passives and modals in the three sub-corpora[13]

5.4 Case study (3): Mossé index for *be*-passives of the six verbs *tell*, *take*, *kill*, *put*, *pay*, *make*

In the third case study the number of *be*-passives across text types was counted for the six verbs *tell*, *take*, *kill*, *put*, *pay* and *make*. Figure 2 gives the number of *be*-passives with these six verbs per 10,000 words in relation to the size of each section.

For all three corpora, creative writing as imaginative prose has a low number of *be*-passives. As for informational prose, both ICE-GB and ICE-NZ have instructional writing as the section with the highest number of *be*-passives; press, academic and non-academic are in medial position. Student essays rank very low in ICE-GB; for ICE-NZ there is a marked difference between student essays and creative writing but student essays too are the category with the second lowest number of *be*-passives. ICE-Fiji lies relatively close in the number of *be*-passives to ICE-GB in academic, non-academic and press. Striking differences to the inner circle varieties, however, can be seen in the usage of the *be*-passive in instructional writing and student essays, which show the reverse position. In ICE-Fiji instructional writing has a particularly low number of *be*-passives and the student essays have a particularly high number of *be*-passives. The low number of *be*-passives in instructional writing in ICE-Fiji seems to correlate with the previous result of showing some preference for *get* in this

category, which may have something to do with the choice of texts or the cultural perception of the text type. As for the high number of *be*-passives in the student essays, it is possible that, in comparison to students who are native speakers of English, those students with an ESL background pay particular attention to the more impersonal style of academic writing they are trying to imitate. This is in line with the results of a Master's thesis on progressives in student essays in ICE-Fiji and ICE-Kenya that found a much higher number of progressive passive forms in this section compared to student essays in ICE-GB (Vogel 2007: 59). On the other hand, lecturers in New Zealand also notice a tendency by New Zealand students to try to be more academic than the academics by using a considerably high number of passives.[14] Although within ICE-NZ the number of *be*-passives in student essays is low in comparison to the other categories, it is remarkably higher than in ICE-GB and close to the amount in ICE-Fiji. This implies that there are different perceptions of formal style at work and that it is not simply an inner circle – outer circle divide.

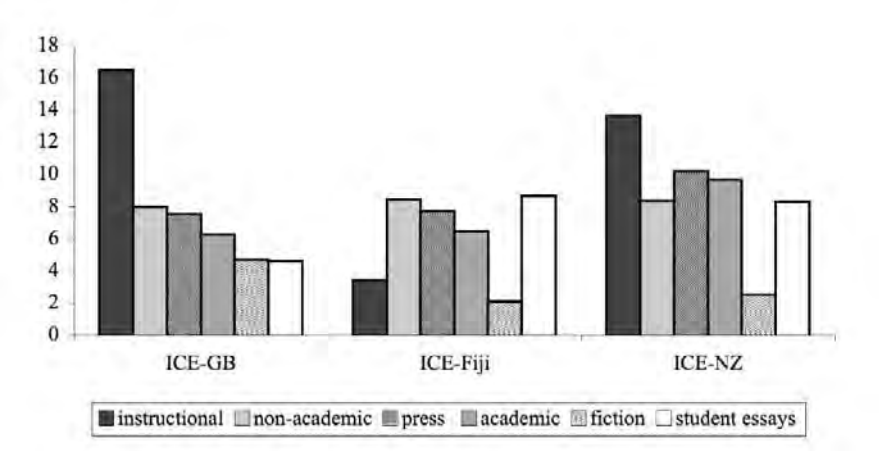

Figure 2: *Be*-passives with *tell*, *take*, *kill*, *put*, *pay* and *make* across text type[15]

6. Conclusion

If one wants to describe passive constructions in Fiji English, one can safely say that comparing the different ICE corpora shows that Fiji English has the same frequency of *be*-passives as inner circle varieties but a lower usage of agentive passives as well as a lower usage of combinations of aspect with voice. As for the usage of *get*-passives, Fiji English uses them less often than New Zealand English but more often than British English. The reasons for these results, however, are a different matter. There is not enough evidence to suggest the influence of a new exonormative model on passive constructions in Fiji English. Rather, stylistic

reasons and a general colloquialisation of the written norm seem to play a role in the usage of the *get*-passive.[16] Socio-cultural differences in the perception of text types in the Pacific in comparison to New Zealand and Great Britain may be a factor too. Fewer agentless passives emerge due to the effects of second language learning. This may be reinforced by the usage of only non-agentive passives in Fijian. Second language acquisition, substrate influence and socio-cultural differences as to the interpretation of text types seem to be crucial in the usage of passive constructions in Fiji English. It could also be seen that corpus design is an important factor to be considered in a corpus-based study on the grammar of Fiji English.

Notes

1 The examples are drawn from interviews I did with Fijians in Suva in spring 2007. Also see Lynch and Mugler (1999: 9, 10) in their discussion of typical features of South Pacific varieties of English as well as Tent and Mugler (2004a: 770, 773).

2 We could also argue that, as English is developing towards a more and more analytic language in which the already restricted inflectional system may be further reduced, number marking of the noun, if a quantifier is used, becomes redundant. The reason then would lie in the typology of the English language. For language internal characteristics of English as another factor to be considered see Sand (2005: 181) and Mair (2003: 85) on the notion of 'angloversals'.

3 For more information see www.state.gov/r/pa/ei/bgn for Fiji.

4 Seventy percent of all mandative subjunctives in her corpus of Fiji English were passive subjunctives whereas for the inner circle varieties the range was between 46 percent and 52 percent of all mandative subjunctives (cf. Biewer forthcoming).

5 The *get*-passive may also be used instead of the *be*-passive to avoid ambiguity with verbs which refer to actions that produce a final result. Does *The chair was broken* refer to the action of breaking or the state of having been broken? *The chair got broken* does not show such ambiguity between static and dynamic meaning (Quirk et al. 1985: 162). Another interesting construction is the passive imperative with *get*, *get vaccinated*, which cannot be replaced by a *be*-passive, as it tells people to arrange for things to be done (Swan 1995: 228).

6 However, the *get*-passive is not likely to become a strong rival of the *be*-passive (Leech et al. forthcoming: 301f). Although the number of *get*-

passives is rising while the number of *be*-passives is declining (Leech and Smith 2006: 194), the overall number of *get*-passives, even in speech, remains comparatively small (cf. discussion in Leech et al. forthcoming: 303).

7 "American English is often assumed to lead in the change towards more informal modes of expression in writing" (Mair 2006: 188) and varieties such as Australian English and New Zealand English are claimed to be more colloquial anyway.

8 *T/A* stands for 'tense/aspect marker', *trs*. stands for 'transitive suffix'. Example from Churchward (1973: 19).

9 Example from Churchward (1973: 20).

10 The suffix *-i* is also used as a transitive suffix before personal pronouns, which is not an ambiguity as the following object makes the function of *-i* clear: *Eratou gunu-vi ko koya*: 'They drink it.'

11 There is no ICE-America so a comparison with this data for American English is not yet possible.

12 And the combination of *by*-agent and progressive was extremely low in the ICE-Fiji component, which further emphasizes that complexity of the form may be an issue here.

13 For absolute frequencies, see table 2a in the appendix.

14 I thank L. Grant and E. Vine for this personal communication.

15 For absolute frequencies, see table 3a in the appendix.

16 We will also have to be careful to distinguish "genuine regional divergence from parallel diachronic developments" (Hundt et al. 2008: 329).

References

Bauer, L. (1994), *Watching English Change. An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London and New York: Longman.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *The Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Biewer, C. (2008a), 'South Pacific Englishes – unity and diversity in the usage of the present perfect', in: T. Nevalainen, I. Taavitsainen, P. Pahta and M. Korhonen (eds.) *Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*. Amsterdam: John Benjamins. 203-219.

Biewer, C. (2008b), 'Concord patterns in South Pacific Englishes – the influence of New Zealand English and the local substrate', in: K. Stierstorfer (ed.) *Anglistentag 2007 Münster. Proceedings*. Trier: Wissenschaftlicher Verlag. 331-343.

Biewer, C. (forthcoming), 'The mandative subjunctive in Fiji English'.

Bolinger, D. (1977), *Meaning and Form*. London: Longman.

Churchward, C.M. (1973), *A New Fijian Grammar*. Suva: Government Press.

Clyne, M. (ed.) (1992), *Pluricentric Languages. Differing Norms in Different Nations*. Berlin and New York: Mouton de Gruyter.

Dixon, R.M.W. (1988), *A Grammar of Boumaa Fijian*. Chicago: University of Chicago Press.

Fischer, S.R. (2002), *A History of the Pacific Islands*. New York: Palgrave.

Geraghty, P., F. Mugler and J. Tent (2006), *Macquarie Dictionary of English in the Fiji Islands*. Sydney: Macquarie Library.

Granger, S. (1983), *The be + Past Participle Construction in Spoken English*. Amsterdam: North Holland.

Greenbaum, S. (1991), *International Corpus of English – The Compilation of the International Corpus of English and its Component, Manual*. London: Survey of English Usage.

Hundt, M. (1998), *New Zealand English Grammar – Fact or Fiction?* Amsterdam and Philadelphia: John Benjamins.

Hundt, M. (forthcoming), 'How often do things get V-ed in Philippine and Singapore English? A case-study on the get-passive in two outer-circle varieties of English'.

Hundt, M. and C. Biewer (2007), 'The dynamics of inner and outer circle varieties in the South Pacific and East Asia', in: M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi. 249-269.

Hundt, M., J. Hay and E. Gordon (2008), 'New Zealand English: morphosyntax', in: K. Burridge and B. Kortmann (eds.) *Varieties of English: The Pacific and Australasia, Handbook of Varieties of English, Volume III*, Berlin: Mouton de Gruyter. 305-340.

Leech, G. and N. Smith (2006), 'Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English', in: A. Renouf, and A. Kehoe (eds.) *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi. 185-204.

Leech, G. and J. Svartvik (1975), *A Communicative Grammar of English*. Longman: London.

Leech, G., M. Hundt, C. Mair and N. Smith (forthcoming), *Change in Contemporary English – a Grammatical Study*. Cambridge: Cambridge University Press.

Leitner, G. (1992), 'English as a pluricentric language', in: M. Clyne (ed.) *Pluricentric Languages. Differing Norms in Different Nations.* Berlin and New York: Mouton de Gruyter. 179-237.

Lynch, J. and F. Mugler (1999), *English in the South Pacific*. On-line publication www.vanuatu.usp.ac.fj/paclangunit/English_South_Pacific.htm, 11.02.05. 1-25.

Mair, C. (2003), 'Kreolismen und verbales Identitätsmanagement im geschriebenen jamaikanischen Englisch', in: E. Vogel, A. Napp and W. Lutterer (eds.) *Zwischen Ausgrenzung und Hybridisierung – Zur Konstruktion von Identitäten aus Kulturwissenschaftlicher Perspektive*. Würzburg: Ergon Verlag. 79-96.

Mair, C. (2006), *Twentieth-century English. History, Variation and Standardization*. Cambridge: Cambridge University Press.

Mangubhai, F. (1984), 'Fiji', in: T.R. Murray and T.N. Postlethwaite (eds.) *Schooling in the Pacific Islands. Colonies in Transition*. Oxford: Pergamon Press. 167-199.

Mangubhai, F. and F. Mugler (2003), 'The language situation in Fiji', *Current Issues in Language Planning*, 4 (3 and 4): 367-458.

Milner, G.B (1990), *Fijian Grammar*. Suva: Government Press.

Nash, W. (1980), *Designs in Prose*. London: Longman.

Odlin, T. (1994), *Language Transfer – Cross-linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.

Poutsma, H. (1926-1929), *A Grammar of Late Modern English*, 2 volumes. Groningen: P. Noordhoff.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language*. London: Longman.

Sand, A. (2005), *Angloversals? Shared Morphosyntactic Features in Contact Varieties of English. Habilitationsschrift*, University of Freiburg.

Schneider, E. (2003), 'The dynamics of New Englishes: from identity construction to dialect birth', *Language*, 79(2): 233-281.

Slewierska, A. (1984), *The Passive – A Comparative Linguistic Analysis*. London: Croom Helm.

Statistics New Zealand (2002), 'Census snapshot: Pacific peoples'. On-line publication www.stats.govt.nz/products-and-services/Articles/census-snpsht-pac-ppls-Jun02.htm, 13.10.2006. 9-11.

Sussex, R. (1982), 'A note on the get-passive construction', *Australian Journal of Linguistics*, 2: 83-92.

Swan, M. (1995), *Practical English Usage*. Oxford: Oxford University Press.

Tent, J. (2001), 'A profile of the Fiji English lexis', *English World Wide*, 22(2): 209-245.

Tent, J. and F. Mugler (2004a), 'Fiji English: morphology and syntax', in: E. Schneider and B. Kortmann, with K. Burridge, R. Mesthrie and C. Upton (eds.) *A Handbook of Varieties of English, Volume II*. Berlin: Mouton de Gruyter. 770-788.

Tent, J. and F. Mugler (2004b), 'Fiji English: phonology', in: E. Schneider, B. Kortmann, with K. Burridge, R. Mesthrie and C. Upton (eds.) *A Handbook of Varieties of English, Volume I*. Berlin: Mouton de Gruyter. 751-779.

Vogel, K. (2007), *Glocalization? A Case-study on the Progressive in Fijian and Kenyan English*. Unpublished thesis (*Staatsexamen*), University of Heidelberg.
Winford, D. (2003), *Introduction to Contact Linguistics*. Oxford: Blackwells.

Appendix

Table 1a: Number of words per section

	instructional	non-ac.	press	academic	fiction	student essays
ICE-GB	43,667	89,140	65,430	88,548	44,692	43,631
ICE-Fiji	23,310	24,890	79,105	91,160	33,750	40,448
ICE-NZ	42,552	86,349	68,866	91,108	43,449	43,368

Table 2a: Absolute frequencies for *by*-agents, progressive passives, perfect passives and modals in the three sub-corpora

	by-agent	prog. passive	perf. passive	modals	all *be*-passives
ICE-GB	112	27	121	169	622
ICE-Fiji	92	20	93	207	767
ICE-NZ	106	25	139	191	724

Table 3a: Absolute frequencies of *be*-passives with *tell*, *take*, *kill*, *put*, *pay* and *make* across text type

	instructional	non-ac.	press	academic	fiction	student essays
ICE-GB	72	71	49	55	21	20
ICE-Fiji	8	21	61	59	7	35
ICE-NZ	58	72	70	88	11	36

Subordinating conjunctions in Middle English and Early Modern English religious writing

Ingvilt Marcoe

University of Cologne

Abstract

This paper presents a contrastive analysis of connective profiles, that is, patterns of distribution of clause-level connectives, with a special focus on subordinating conjunctions in Middle English and Early Modern English religious treatises and prayers. It will be shown that differences in the connective profiles of the two genres may point to genre-specific functional motivations, whereas similarities in the profiles may not necessarily indicate uniform causes for the use of particular connectives. Instead, the similarities may actually hide different functional motivations. Additionally, diachronic changes in the distribution of clausal connectives are, it will be argued, genre-specific to the extent that they are triggered by changes in genre conventions.

1. Introduction

The history of clause-level connectives has attracted increased attention in recent corpus-based studies. Apart from studies with a special focus on the diachronic development of individual connectives (Rissanen 2007; Molencki 2007; Sorva 2007) or on connectives expressing particular types of semantic relations (Lenker 2007; Claridge 2007), another approach has been the analysis of co-occurrence patterns of clause-level connectives (Kohnen 2007). Kohnen identifies distinct patterns of distribution in the major clausal connectives, so-called "connective profiles", in the diachronic development of sermons and the synchronic comparison of sermons and statutes.

In this pilot study, the same approach will be pursued further with the aim of gaining new insights into the distribution of clausal connectives across genres. The distribution of subordinating conjunctions in the connective profiles of late Middle English (ME) and Early Modern English (EModE) religious treatises and prayers will be contrasted. To explain similarities and differences in the distribution of these connectives in the two genres, functional motivations which influence the choices writers make regarding the means of expression at their disposal will be taken into consideration. By establishing connective profiles in a variety of genres across time, diachronic changes in the use of connectives may also be traced in more detail. On the one hand, general trends, which are not limited to single genres, may be identified. Such general trends mentioned in the

literature include, for instance, the decline of subordinative *for* in favor of the causal subordinator *because* in the development of the English language (see Rissanen 1998, 1999; Claridge and Walker 2001) or the genesis of adverbial subordinators expressing cause, condition and concession from subordinators with locative, temporal and modal senses, which Kortmann describes as a diachronic universal taking place across various languages (1997a and 1997b). On the other hand, the comparison of connective profiles may also bring to light changes that can be interpreted as genre-specific developments, which reflect changes in genre conventions and altered communicative needs of genre users.

2. Categories of analysis

For the present analysis, the focus is on the distribution of subordinating conjunctions on the clause level. These have been classified according to the functional and semantic types of clauses they mark. Thus, an initial distinction is made between subordinators marking nominal clauses and subordinators introducing various semantic types of adverbial clauses.

Subordinators introducing nominal clauses are further differentiated into *that*-, zero *that*- and *wh*-clauses (see examples (1) and (2)). Although no explicit subordinator is present in zero *that*-clauses, they have been considered in the present analysis since the subordinator *that* may potentially be inserted into the complex sentence (Quirk et al. 1985: 1007). Nominal clauses which are introduced by a *wh*-element constitute *wh*-interrogative, yes/no and alternative interrogative, exclamative and nominal relative clauses (see Quirk et al. 1985: 1050-1061 for more detail).

(1) Any reasonable man would think *Ø* those men have not really such an abhorrence of Popery as they pretend, and *that* there might easily be found terms of accommodation between them. (Goodman, *Compassionate Inquiry*, 1674, 15.15-19)

(2) Remember not, O Lord, *how* we have been full of envy and malice, anger and revenge, fierce and earnest in the purchases and vanities of the world. (Taylor, *Golden Grove*, 1654, 100.18-21)

Adverbial subordinators are classified according to the semantic relation expressed between matrix and subordinate clause. Since there is neither a uniform way of classifying adverbial clauses, nor a fixed terminology, the categories of analysis used in the present study will be discussed briefly. They are based on the classifications of adverbial clauses found in the standard literature, but are slightly modified for the purpose of compression (see Quirk et al. 1985: 998, 1077-1112; Biber et al. 1999: 776-853; as well as Rissanen 1999; for lists of subordinators see Kortmann 1997b: 116).

First, clauses of time and place have been combined into one category expressing all spacio-temporal relations between subordinate and matrix clause.

Subordinators marking clauses of time and place are, for instance, *when*, *while*, *whilst*, *since*, *sith*, *after*, *before*, *ere*, *until*, *now that*, *against*, and *as soon as*, for example:

(3) And all our glory / ioye & felicite / shalbe *whan* in perfet loue and fruicion / we shal se the father of heuen… (Bonde, *Pilgrimage of Perfeccion*, 1526, 4.31-32)

The second category consists of adverbial clauses of manner which give a description of how the action described in the matrix clause is carried out. These are introduced by the subordinators *as*, *as though*, and *as if*. Often, the manner of one state or action is compared to the manner of another state or action (example (4)).

(4) Peter .j. Pet .iij. exhorteth wives to be in subieccion vnto their husbands […] *as* Sara obeyed Abraham and called him lorde. (Tyndale, *Obedience*, 1528, int6.10-15)

The semantic categories of condition, concession, purpose, result, and reason are subsumed under the more comprehensive category of contingency in Biber et al. (1999). However, since Biber et al.'s corpus-based analysis revealed that finite clauses fall predominantly into this category (787), a more detailed distinction has been drawn in the present study. While Quirk et al. (1985) treat these types of adverbial clauses individually, Rissanen (1999) takes an intermediate approach, which will be followed here.

Clauses of condition and concession have been grouped together since there is some semantic overlap. As Rissanen (1999) points out, a condition is involved in both clause types. In conditional clauses, the action in the main clause depends on the fulfilment of the condition in the subordinate clause. In concessive clauses, a negative condition is expressed: the action in the main clause does not depend on the fulfilment of the condition in the subordinate clause or is even contrary to it (see example (5) for both instances).

(5) And servauntes obeye your masters with all feare not only *if* they be good & courteous: but also *though* they be frowarde. (Tyndale, *Obedience*, 1528, int7, 10-20)

The semantic overlap becomes even more evident when considering that, in ME, the now typically conditional subordinator *if* could be used in concessive clauses, while the now typically concessive subordinator *though* could also occur in conditional clauses (Rissanen 1999: 307; Quirk et al. 1985: 1088). Other subordinators introducing conditional and concessive clauses are, e.g., *although*, *albeit*, *howbeit*, *whereas*, *notwithstanding (that)*.

Clauses of result and clauses of purpose have also been combined since they express similar meanings. The major difference is that clauses of purpose

indicate results which are putative, meaning that the results are intended but have not been achieved up to that point, whereas resultative clauses describe factive results, which have already been obtained in the past. In addition, the two clause types may be marked by the same subordinators, namely *as* and *that*. These conjunctions are often, but not necessarily, introduced by emphasizers in the main clause, particularly by *so* and *such*, but also by prepositional phrases such as *to the purpose (that)* and *to the end (that)*. Negative clauses of purpose or result may also be marked by the subordinator *lest*, as in example (6) (see Rissanen 1999: 304; Quirk et al. 1985: 1107-1108).

(6) Take from me a wanton eie, and let me not gaze vpon the beautie or comlie personage of anie man, *least* I be taken in his snares. (Bentley, *Fifth Lampe*, 1582, 2.19-21)

Clauses of reason express a causal relation between the two propositions expressed in matrix and subordinate clause. They are made explicit by subordinators such as *because*, *by cause (that)*, *as*, *forasmuch as*, *since*, and *for (that)*. With regard to the conjunction *for*, it is important to note that the distinction between coordinators and subordinators is not absolutely clear-cut. In Present-day English, this conjunction maintains an in-between status, being considered neither a pure coordinating nor a pure subordinating conjunction. This is due to the fact that the connective *for* lacks some features typical of subordinators and others typical of coordinators (see Quirk et al. 1985: 927 for a more detailed discussion). From a diachronic perspective however, the syntactic status of *for* has always been rather ambiguous, shifting from the more subordinative towards the more coordinative end of the cline. According to Jucker (1991: 218), the use of *for* changed from predominantly subordinative to predominantly coordinative contexts after the ME period. Similarly, Rissanen points out that "*for* lost its syntactic subordinator features in Early Modern English" (1989: 3). In ME, *for* may still occur in subordinative contexts, i.e. in front of the main clause or as part of a series of coordinated causal clauses (see example (7)). In EModE, it favors coordinative contexts, i.e. post-position with a loose causal link to the preceding clause (Rissanen 1999: 306) (see example (8)).

(7) And if þou haue sorowe for þi sinnes, & *for* þou ert swa lang in exile, owte of þi contre, & forsakes þe solace of þis lyfe: þou sal haue for þis sorow, þe ioy of heuen. (Rolle, *Forme of Liuing*, a1349, 41.20-25)

(8) Christes commaundemente is not regarded: *for* nothing is hearde commonly in the Churche but in a straunge tongue that the people doth nothing vnderstande. (Ridley, *Pituous Lamentation*, 1566, int4.15-29)

The category 'others' contains less common semantic types of adverbial clauses, such as clauses of exception (*but that*, *except that*), proportion (*as…so*), extent and degree (*insofar as/that*, *insomuch as*), and preference (*rather than*) (see Quirk et. al 1985: 1102-1112).

Of course some subordinators overlap in meaning. Subordinators are generally assumed to have one central and one or more peripheral meanings. For example, the subordinator *as* can introduce four different types of adverbial clauses, i.e. clauses of time, clauses of manner, clauses of reason, as well as clauses of proportion. Likewise, the subordinator *while* may introduce both temporal and concessive clauses. As a result, the semantic relations expressed by particular subordinators may at times be ambiguous. However, the meaning can usually be deduced from the context; when more than one reading is possible, the primary meaning of the subordinator is generally assumed (see Rissanen 1999: 302; Biber et al. 1999: 787; Quirk et al. 1985: 1077).

Elements which constitute a link on the phrase level will be excluded from the present analysis. As a result, relative pronouns have not been considered since relative clauses mostly serve as postmodifiers of noun phrases, with the relative pronoun linking a clause to a phrase. Additionally, non-finite clauses have been ignored as they are not typically marked by an explicit subordinator, being considered subordinate due to the absence of a finite verb.[1]

The study is based on the analysis of a corpus of 202,194 words, consisting of religious treatises and prayers (a complete list of texts is given in the appendix). These have been further sub-divided according to time periods: ME (fourteenth and fifteenth centuries) and EModE (sixteenth and seventeenth centuries). Since the sub-corpora are rather small in size, the conclusions drawn from the present study may only be regarded as tentative. Although the fact that the analysis deals with relatively high frequency items somewhat makes up for the small sub-corpora, further studies need to be conducted to confirm the results presented here.

3. Results

When looking at the distribution of subordinators in religious treatises and prayers (see tables 1 and 2), the use of subordinating conjunctions does not seem to differ greatly at first sight. Except for the fact that the overall frequencies of subordinators are much lower in prayers, the proportion of subordinators introducing nominal and adverbial subordinate clauses is very similar in both genres (ratio of about 1:2 nominal to adverbial clauses) with a slight overall increase in the EModE period.

A closer look at the distribution of adverbial and nominal subordinators in the two genres will offer some valuable additional insights, since it remains to be seen whether this similarity in the connective profiles can be upheld in a more detailed analysis of these types of subordinators. Therefore, it will be interesting to see in how far the distribution of subordinators marking various types of adverbial and nominal clauses is uniform or different in the two genres. To explain such similarities or differences, an analysis of the motivations shared by the genre users follows.

Table 1: Distribution of nominal and adverbial subordinators in religious treatises (absolute numbers and relative frequencies per 1,000 words)

	ME 64,761 words		EModE 62,743 words	
Adverbial subordinators	1226	18.9	1203	19.2
Nominal subordinators	633	9.8	720	11.5
Total subordinators	1859	28.7	1923	30.7

Table 2: Distribution of nominal and adverbial subordinators in prayers (absolute numbers and relative frequencies per 1,000 words)

	ME 13,581 words		EModE 61,109 words	
Adverbial subordinators	139	10.2	852	13.9
Nominal subordinators	77	5.7	460	7.5
Total subordinators	216	15.9	1312	21.5

3.1 Distribution of adverbial subordinators in the texts

The detailed analysis of adverbial subordinators (see tables 3 and 4)[2] shows that both similarities and differences become visible in the connective profiles of the two genres. Most notable is, firstly, the difference in the overall distribution of connectives: while it is quite varied in religious treatises, it seems rather restricted in prayers, especially in the EModE period, in which only one type of subordinator occurs with a relative frequency higher than 2.0. A similarity shared by both genres is the increase in conditional, concessive and causal subordinators combined with a decrease in subordinators marking clauses of time and place. The distribution of clauses of purpose and result differs again in the two genres, in that they decrease slightly in treatises while they increase quite dramatically in prayers.

The similarity in the distribution of adverbial subordinators, namely the increase in conditional, concessive and causal subordinators, which coincides with a decrease in subordinators in clauses of time and place in both genres, can be related to the results of Kortmann's (1997b) diachronic cross-linguistic study of adverbial subordinators. Kortmann follows a typological approach to language change. By comparing linguistic changes across different languages, he attempts to discover universal paths of language change, "diachronic universals", which are described as gradual, regular, and uni-directional (1997b: 111). Kortmann's findings indicate uni-directionality of semantic change with regard to certain adverbial subordinators. He postulates that subordinators expressing so-called

"CCC-senses" (i.e. condition, concession, cause or a related sense such as purpose or result) generally develop from temporal, locative and modal senses (e.g. *as*, *if*, *when*, *whereas*, *while or whilst*) and points out that the Present-day English inventory of temporal subordinators has become smaller, whereas the subordinator inventories expressing CCC-relations have expanded. Therefore, Kortmann speaks of a "dramatic change" in the semantic composition of subordinator inventories in the history of English. By comparing these findings regarding the English language to results obtained in the study of other European languages, Kortmann assumes a link between the development of a written standard and the increase in subordinators expressing CCC-relations. He argues that CCC relations are important for establishing discourse coherence in written language, especially in argumentative writing.

Table 3: Distribution of adverbial subordinators in religious treatises according to clause types they mark (absolute numbers and relative frequencies per 1,000 words)

	ME 64,761 words		EModE 62,743 words	
Subordinators in:				
time/place clauses	346	5.3	281	4.5
manner clauses	242	3.7	222	3.5
conditional/concessive clauses	307	4.7	354	5.6
purpose/result clauses	217	3.4	167	2.7
cause/reason clauses	62	1.0	154	2.5

Table 4: Distribution of adverbial subordinators in prayers according to clause types they mark (absolute numbers and relative frequencies per 1,000 words)

	ME 13,581 words		EModE 61,109 words	
Subordinators in:				
time/place clauses	42	3.1	100	1.6
manner clauses	17	1.3	89	1.5
conditional/concessive clauses	13	1.0	117	1.9
purpose/result clauses	58	4.3	458	7.5
cause/reason clauses	8	0.6	66	1.1

An extension of the inventories of subordinators expressing CCC-relations may be tied to an increased need on the part of speakers to express these types of

semantic relations in different discourse situations. This would be reflected in heightened frequencies of these clause types, as found in the present study. However, the present study takes a more pragmatic rather than a typological approach. The focus here is not on the study of diachronic universals but rather on diachronic variation of language use in context. Since the communicative intentions of people affect the choices they make in their use of language, variation in the way language is used in texts differentiated by communicative purpose, discourse situation and subject matter is taken into consideration here. A look at the way writers with similar communicative purposes employ adverbial subordinators can help elucidate reasons and motivations for the choices they make. As Rissanen (1996: 277) points out, the study of variation and change differentiated by genres constitutes an attempt to describe people's purposes of communication and to relate these to the socio-cultural and political conditions of the speech community. Changes in the socio-historical background of genre users can thus be seen to have an effect on the genre conventions of the time and therefore also on the writers' choices regarding the linguistic means of expression available to them. As a consequence, the rise in conditional, concessive and causal clauses as found in religious treatises and prayers in the present study may be interpreted, along the lines of Kortmann, as part of a general development, that is, a diachronic universal. The similar development occurring in genres such as prayers and religious treatises – the former being designed to be performed orally, and the latter constituting a typically written genre – clearly points in this direction. However, when looking at the developments in their contexts of occurrence, changes in genre conventions and socio-cultural developments seem to have additionally influenced the writers' choices in their use of subordinators expressing CCC-relations.

With regard to prayers, it would be difficult to show such an influence with certainty, since the frequencies for conditional, concessive, and causal clauses remain relatively low and increase only slightly. Prayers appear to be a rather stable genre with little diachronic change in the distribution of clausal connectives and limited variation in the use of different subordinators. This stability results from the relatively fixed communicative setting and purpose of prayers across time. They constitute addresses to a transcendental being which are designed for the expression of requests, admission of guilt and sins, for thanks-giving and praise. Moreover, they are intended for public service or private prayer, can be performed individually or in groups, and they are often relatively short and therefore published in collections of prefabricated texts (see Kohnen forthcoming).

Considering religious treatises, changes in the communicative setting and purpose may be seen to influence the distribution of conditional, concessive, and causal connectives. The increase in conditional and concessive subordinators in religious treatises will be considered first, by looking at the use of these subordinators in their typical contexts of occurrence. In EModE religious treatises, conditional subordinators are increasingly used in passages of text which serve an argumentative function. The use of conditionals by EModE

pamphleteers for argumentative purposes has already been commented on by Claridge (2007). According to Claridge (2007: 252), conditionals are used in political and scientific pamphlets to present options and alternatives. They may additionally be used to integrate the readers into the conclusion-drawing process by asking them to draw their own conclusions concerning a particular topic. The use of conditionals in EModE religious treatises serves argumentative purposes as well: here conditionals are employed to describe courses of action as conditions for ensuing positive or negative events, i.e. for benefits or sanctions. In this way, clauses of condition and concession occur in sections of the text in which the author attempts to persuade the reader to follow a particular course of action. This is the case in example (9), in which the reader is presented with alternative courses of action and their ensuing consequences. Additionally, clauses of condition serve as a means of strengthening the validity of the author's arguments by including the reader in the author's thought process. In example (10), the reader is addressed directly by means of the second-person pronoun *you* as well as the first-person plural inclusive *we* to motivate him or her to take an active part in the considerations.

(9) *Iff* thou obeye (though it be but carnally (eyther for feare / for vayne glorie or profit) thy blessinge shalbe longe life vppon the erth. Contrary wise *if* thou disobeye them / thy life shalbe shortned vpon the erth. (Tyndale, *Obedience*, 1528, int2.13-17)

(10) *If* we saw a man killed and cut in pieces by the way, we would presently ask, Oh who did this cruel deed? *If* the town were wilfully set on fire, you would ask, What wicked wretch did this? So when we read that the most will be fire-brands of hell forever, we must needs think with our selves, How come this to pass? (Baxter, *Call to the Unconverted*, 1658, 2.22-3.6)

As noted above, the use of causal subordinators also increases in EModE religious treatises. Causal clauses may be sub-categorized according to the causal relations they express. One major distinction is commonly drawn between direct and indirect reason clauses, with the latter type expressing the reason for the speech act expressed in the matrix clause. Clauses indicating direct reason relationships can moreover be differentiated into external and internal reason clauses (see, e.g., Quirk et al. 1985: 1103-1107; Rissanen 1998: 393; Lenker 2007: 198). This distinction is particularly relevant for the choice of causal subordinators in religious treatises.

External reason clauses express the author's perception of an inherent objective connection between states or events in the real world. Thus, the cause is described as being based in external reality. They are typically used to make justifications for particular claims, as the author states causal relations as facts which are based in the world around him. In example (11), the behavior of God is justified in terms of his undeniable physical and mental attributes (i.e. his almighty power and unquestionable will); in example (12), reasons why one should never trust the devil are based on the devil's inherent qualities of being an

enemy and a liar. In the print version of the text, the bible reference John 8:44 is given in the margin as additional evidence for the validity of the claim.

(11) God might tourne the will of the wicked into good, *because that* he is almightie, plainly, he could doe it, wherefore then doeth he it not? *Because that* he will not, why he wilt not, that doe we leaue vnto him, for we ought not to be wiser then it behoueth. (Northbrooke, *Poor Mans Garden*, 1571, 7.20-26)

(12) But speake the truth, is it not rather a strong temptation of Sathan your deadly enemie to trouble the peace of your conscience, and if it be possible, to driue you to desperation. If it be so, as I feare greatly, then I say unto you, there is no cause why you should beleeue him. First *because* he is a lyer. Secondly *because* he is your enemy, who meanes you no good at all. That he is a lyar it is manifest, because he hath beene so from the beginning, And he cannot nowe chaunge his nature. (Linaker, *Comfortable Treatise*, 1595, 17.3-15)

Internal reason clauses express an author's inference of a connection, in other words, his assumptions regarding the reasons for subsequent conditions, states or events, which are based on a deduction process from facts known to him. Since these reason-consequence relations are based in the writer's world of reasoning, they are more subjective in nature. In EModE texts, internal reason clauses occur in passages in which the author wants to argue a point by making assumptions concerning the reasons for certain claims. In the following example, the first two causal clauses introduced by "because" constitute internal reason clauses. The author infers possible reasons for a claim, which he is actually in the process of refuting. The reasons are therefore purely hypothetical and not based on a connection in the real world. Tillotson indicates that he is only inferring a possible (but to him not likely) connection by using epistemic *must* in the matrix clause ("it must be from one of these two reasons").

(13) They pretend for this Doctrine [of Transubstantiation] the Authority of Scripture in those words of our Saviour, This is my body. Now to shew the insufficiency of this pretence, I shall endeavour to make good these two things 1. That there is no necessity of understanding those words of our Saviour in the sense of Transubstantiation. If there be any, it must be from one of these two reasons. Either *because* there are no figurative expressions in Scripture, which I think no man ever yet said: or else *because* a Sacrament admits of no figures; which would be absurd for any man to say, since it is of the very nature of a Sacrament to represent and exhibit some invisible grace and benefit by an outward sign and figure... (Tillotson, *Discourse against Transubstantiation*, 1684, 4.6-27)

The increase in external reason clauses reflects an increased desire on the part of EModE writers to give explanations and justifications for their claims. The use of

internal reason clauses shows that these writers also felt a greater need to express inferred cause-and-effect relations, especially in more elaborate discussions of religious matters. In addition, the rise in conditional and concessive clauses used in argumentative passages reflects the writers' extended need to persuade readers of a particular point of view. These marked changes in the choices EModE authors made in their use of language are clearly initiated by changes in genre conventions and socio-cultural changes occurring in the religious discourse world in the EModE period. ME treatises were still predominantly expository in nature. Their focus on the presentation of the basic tenets of the Christian faith can be seen as a result of the Fourth Lateran Council (1215), which decreed the duty of yearly confession for all Christians. Hence, knowledge of the fundamentals of the faith had to be passed on, as a first step, to the lower clergy and, from there, to the largely illiterate and unlearned laity (Barratt 1986: 413). Therefore, ME treatises often represent handbooks for lower clerics, which are intended as aids for the instruction of parishioners. They often comprise expositions of individual points of faith (e.g. *Exposition of the Pater Noster*, *Litil Tretis on Seven Deadly Sins*), guidebooks with advice on the practical application of the doctrine to everyday life (e.g. *Speculum Christiani*), and manuals on preparation for death (e.g. *Craft of Dying*). However, in the EModE period, a shift takes place in the writing conventions of religious treatises. The texts become increasingly elaborate and argumentative in style. The presentation of the basics of the doctrine is increasingly supplemented by detailed discussions of more advanced points of Christian belief. With the break of the Anglican Church with Rome, religious treatises start to address more controversial religious topics, dealing with divergent Catholic and Protestant, or inter-Protestant views on religious matters such as royal versus papal supremacy, transubstantiation, and predestination (Tyndale's *Obedience of a Christian Man*; Tillotson's *Discourse on Transubstantiation*) (see also King 2000: 115-120). In religious texts of the EModE period, it is no longer sufficient to simply state the principles of the Christian doctrine to instruct the reader in the basic tenets of Christian faith. With the diversification of beliefs, writers feel compelled to argue their points of view and to add justifications for the beliefs they wish to disseminate.[3]

A point which further underlines the genre-specific distribution of adverbial subordinators is the divergent diachronic development in the distribution of purposive and resultative subordinators in religious treatises and prayers (see tables 3 and 4 above). There is an increase in prayers, and a decrease in religious treatises. Although the decrease in religious treatises is fairly small and should therefore be considered with care, a comparison of the contexts in which purposive and resultative subordinators occur in both genres still shows interesting differences regarding the motivations for their use.

In ME treatises, clauses of purpose and result are often used to give recommendations on how to lead a pious life. In these instances, clauses of purpose and result serve as a means of exhorting the readers; the author urges them to act in a certain way and depicts, almost as an incentive, a positive outcome (see examples (14)-(15)).

(14) Wherfore, if þere come sumtyme temptaciouns and tribulaciouns þe whiche ben ordeyned for to ponesche and for to clense Goddis children, and deuocioun be wiþdrawe, strenkþe þee þanne not þe lesse for to praie, ne to wake, ne to faste, and not to þe lesse in oþere gode werkis for to stonde. *So þat* þoru contynuaunce in praieris wiþ þe teris of þin iȝen unceessably excercisen þee, þat þou myȝt, as it were, constreyne God to ȝeue þee feruour and heete of holy deuocion. (Hilton, *Perfection*, ?a1396, 3.12-4.6)

(15) Qwhils þou ettis and drynkis, mynde of þi god þat þe fedis fro þi mynde pass not, bot prais, blys & glorify hym in ilka morsel, *so þat* þi hart be more in goddis louynge þen in þi meet, *þat* þi saule fro god be not partyd be any howr. (Mysin, *Mending of Life*, 1434, 113.21-24)

Clauses of purpose and result may additionally be used to describe the intentions or outcomes of certain actions or events when instructing the reader in the Christian doctrine. As discussed above, this function is more prominent in ME texts, but still persists well into the EModE period and reflects the genre's function of religious instruction. In example (16), the 'Fall of Man', and in example (17) the 'Passion of Christ', is narrated.

(16) Aftire þe tyme þat man was exilet oute of þe hie cite of heuen by þe riȝtwis dome of almiȝty god souereyn kyng þerof, for his trespasse & his synne, & so wrecchedly lay in prison & was halden in þe handes of þat Tyrant þe deuel of helle, *þat* none miȝt come aȝeyne to þat blessed Cite þe space of fyue þowsande ȝere & more, alle þe blessede spirites of heuene […] hadden grete compassion of so longe meschefe of manne. (Love, *Mirror*, a1410, 13.38-14.6)

(17) Hee was deliuered to death for our sinnes, as if he should say, whatsoeuer griefe or torment hee endured liuing, or dying, hee endured it for our sakes, *that* the whole fruit & comfort therof might redound [return] to vs. (Linaker, *Comfortable Treatise*, 1595, 7.21-26)

In EModE treatises, however, recommendations for pious living and pure exposition of the doctrine begin to play a less prominent role, a development which is reflected in the slight decline of this clause type during the EModE period. The motivations for the use of purposive and resultative clauses change: they are now increasingly found in argumentative passages, in which the author attempts to influence the readers in their thought processes. In example (18), the author invokes the image of an angel who speaks to the readers, then describing the hypothetical consequences of such a vision as a means of convincing the readers of their need to convert to God; excerpt (19) is taken from a passage in which the author attempts to persuade the readers of the insurmountable differences between the Catholic and Protestant faiths.

(18) If you had but once heard this word by the voice of an Angel, Thou must be Converted or Condemned; Turn or Die: would it not stick in your mind, and haunt you night and day, *so that* in your sinning you would remember it, and at your labour you would remember it, as if the voice were still in your ears, Turn or Die. (Baxter, *Call to the Unconverted*, 1658, 24.4-14)

(19) Or how comes it to pass that all those of the Roman Communion withdraw themselves from ours, and are commanded so to do by the Head of their Church under peril of damnation? And on the other side the true Protestants of the Church of England, think it their duty to absent themselves from Roman Worship, *lest* they should defile their Consciences with their Superstitions? I say how comes this distance and apprehension of sin and danger reciprocally, if the differences between them be inconsiderable? (Goodman, *Compassionate Inquiry*, 1674, 14.15-26)

Coming now to subordinators introducing clauses of purpose and result in prayers, the data in table 4 reveal that these subordinators show the highest relative frequencies among all types of adverbial subordinators and increase further in EModE. Interestingly, this clause type occurs mostly in combination with matrix clauses in which a request is expressed to God or Christ, or sometimes to the Virgin Mary, an angel, or a saint. These adverbial clauses are typically used to justify the utterance of the demands by indicating the intentions and purposes behind them. The question which poses itself is: to whom are these justifications addressed – to the primary addressee, i.e. the transcendental being, who is also the addressee of the request, or to a secondary addressee, which is the Christian community that perform these prayers? If the latter were the case, then the indication of the purpose of the request to God may be designed to provide additional information to the people praying as a means of instructing them in the fundamental principles of their faith. This may even explain the increase in clauses of purpose and result in prayers in the EModE period: with the possibility of publishing prayers in fairly cheap collections with a wide circulation, prefabricated prayers could reach a broader audience with a limited knowledge of religious matters (see Kohnen forthcoming).

Nonetheless, the primary addressee should not be discarded so quickly. According to Kohnen in his discussion of the frequent use of relative clauses and appositions in prayers (forthcoming), additional information does not necessarily represent unshared knowledge between addressor and addressee. Additional information may also represent 'unused' information. This unused information constitutes shared knowledge between addressor and addressee which has not yet been used in the communicative context but which is somehow relevant to the point being made (see also Prince 1981). When making demands, it is often necessary to negotiate the general terms, i.e. to express in how far the outcome of the demand may be favorable to both parties. When looking at the purposes for the demands expressed in the prayers, it becomes clear that they involve a behavior or state on the part of the addressor which is in accordance with the Christian doctrine or which serves as praise to God or Christ (see examples (20)-

(22)). The purposes for such demands would therefore be favorable to both parties and thus constitute valuable arguments in the negotiation process.

(20) Fyl my herte with pyte, compassion & mercy / *that* I may gyue forth liberally vnto all men / especially to my enmyes the benefitis of perfit loue with pure affecte. (anon., *Mystic Rosary*, 1533, 21.7-11)

(21) I therefore most humblie beseech thee, O mercifull Father […] that thou wilt so order my tongue, and dispose my talke, *that* I speake nothing, but that becommeth my state, age, and person; neither *that* I delight to heare anie talke, that might in anie point mooue me to lewdness or lightness, seeing that euill words corrupt good manners. (Bentley, *Fifth Lampe*, 1582, 1.7-11)

(22) Sanctifie me (O holy Father) this holy day, with thine especiall grace, that I may honour thee as a Creator, loue thee as a redeemer, & expect thee as a Sauiour, and that I may haue modest carriage in my behaviour, true deuotion in my Prayers, and reuerent attention to heare thy sacred word; and so vnlocke the eares of mine vnderstanding, *that* I may obserue, learne and imbrace, such things as a necessary for mee, to the better confirmation of my faith in Christ Iesus, and the saluation of my soule by his blood. (Sorocold, *Supplications*, 1612, 4.5-6.1)

Clauses of purpose and result are additionally used to describe the purpose of Christ's actions on earth, as in example (23), or to describe the result of actions committed against Christ (example (24)). One could argue again that these clauses are intended to give additional information to the persons performing these prayers. They may, however, also be seen as part of the above mentioned negotiation process between the person praying and God, since the descriptions of Christ's Passion serve as arguments for why the demands of the addressor should be heard and consequently granted.

(23) O Lord my God, accept my prayers; O good Jesus, Saviour of the world, which gavest thy self to the death of the Cross, *that* thou mightest save sinners, regard me a wretched sinner, calling upon thy name; and take not heed so to my wickedness, that thou forget thy goodness. (Church of England, *Primer*, 1658, 223.10-16)

(24) O blessid Ihu maker of al the worlde that of a man may not me mesured / whiche closest in thy honde all the erthe. Haue mynde of thy bitter sorow. Firste whan the Iewes fastened thy blessid hondes on the cros. wyth blunt nayles […] And soo cruelly they drew thy blessid body in lengthe and brede to the mesure of the cross *that* all thy Ioyntes of thy limmes were both losed and forbroken. For mynde of thy blessed passion / I beseche the benygne Ihesu giue me grace to keepe wyth me bothe thy loue and thy drede / (anon., *Fifteen Oes*, 1491, int3.5-21)

Finally, we need to ask the question of why clauses of purpose and result are less common in ME prayers. In this case it is difficult to say whether the lower frequency is due to the idiosyncratic style of individual authors or whether it is a general characteristic of early English prayers, since the number of extant ME prayers is, unfortunately, limited. In the present ME data, fewer requests are made and the texts contain longer passages of praise to God (see example (25)). ME prayers thus seem to be more contemplative in nature.

(25) My lorde God, I vnderstone well that ye of your grace hath made me of no thynge and giuen me beyng amonge your creatures, and truly whan I was noo thynge I myghte noo thyng deserve. Thenne all this that I am & haue I haue received of your specyall gyfte & grace, wythout my deserte. And of your creatures there ben some hyer / some lower. And I know well that ye myghte haaue made me the most vyle creature that is, and this dyde ye not, But of your bountee formed me to be amonge the mosste hie creatures tha is, this is to knowe, aungell & man that in your likeness shall see you in your glorye. And this dignyte haue ye gyuen me wythoute my deserte, yf I lese it not by my defawte. And by thys reason, mercyful lorde, am I enteerly bounde soueraynly to loue you wyth all my soule, wyth all my herte, and wyth all my power. (anon., *Prayers Love*, a1500, 104.7-20)

As the above discussion has shown, the divergent developments regarding clauses of purpose and result in both genres may very well be explained by considering their communicative functions in the texts, as well as overall changes in genre conventions. In prayers, their increasingly common use is closely linked to the overall aim of prayers to make requests to God. In treatises, on the other hand, they are predominantly used in the early texts as a means of instructing the reader in matters of faith, for instance, to present the consequences of moral or immoral behaviour, or they are used in descriptions of Christ's Passion. Their frequency decreases in later texts, in which the exposition of religious points becomes more elaborate and argumentative.

3.2 Distribution of nominal subordinators in the texts

Tables 1 and 2 above revealed that the ratio of nominal to adverbial subordinators is similar in both genres (with a ratio of 1:2). Does this imply that the functional motivations for the use of nominal subordinators are also alike in both genres? Nominal clauses function predominantly as complements to verbs. For an investigation of the reasons for the use of subordinators introducing nominal clauses, *that*- and zero *that*-clauses functioning as complements to verbs have been analyzed in more detail. Different semantic types of 'governing' verbs in the matrix clause may be distinguished. The classification found in the *Longman Grammar of Spoken and Written English*, in which three types of verbs are distinguished ('mental verbs', e.g. *think*, *believe*; 'speech act verbs', e.g. *say*, *declare*; 'other communication verbs', e.g. *show*, *suggest*) seems most useful, since it allows one to differentiate between verbs introducing indirect speech and

verbs expressing intellectual states (Biber et al. 1999: 662-666).[4] In the present analysis, the latter two verb types have been combined into a single category for the purpose of simplification; thus, a distinction is drawn between mental verbs (e.g. *think*, *believe*) and communication verbs (e.g. *say*, *declare*, *show*, *suggest*), which are additionally differentiated according to concord with first-, second-, and third-person subjects.

Table 5: Distribution of verbal constructions controlling nominal (zero) *that*-clauses which function as complements to verbs in religious treatises (absolute numbers and relative frequencies per 10,000 words)

	ME 64,761 words		EModE 62,743 words	
Subordinators with:				
1st person mental verbs	22	3.4	29	4.6
2nd person mental verbs	56	8.6	37	5.9
3rd person mental verbs	72	11.1	82	13.1
1st person communication verbs	25	3.9	35	6.5
2nd person communication verbs	9	2.0	14	2.2
3rd person communication verbs	80	12.4	165	26.3

The figures in table 5 show the distribution of verbal constructions governing *that*-clauses in religious treatises. Mental and communication verbs with third-person subjects show the highest relative frequencies and increase further in EModE. First-person mental and communication verbs also increase in EModE, while second-person mental verbs decrease. Since the main purpose of religious treatises is to instruct the reader in religious matters by presenting or discussing the main principles of religious doctrine,[5] it is not surprising to see that the most common types of verbal constructions occur either in expository passages, in which the doctrine is expounded, or in argumentative sections, in which religious subject matters are discussed in an argumentative way.

In expository sections of religious treatises, communication verbs with third-person subjects are frequently used to introduce reported speech. They serve to quote the statements of others and, especially in ME texts, to make references to Church authorities in order to validate statements or claims made by the author. This use is nicely illustrated by the following two excerpts:

(26) Tresown is whan a man is fals to hym. to whom he owȝte to be trewe for wynnyng of money or getyng of mede. In þis bronche of couetyse trespacyd Judas […] as it is rehersid in the gospel. Mt 26. And *þe holy man Bede seyth þat* his folweris be alle þo men þat for eny meede beryn fals witnesse in doom a ȝens her euyncristyn as þe lawe of holy cherche rehersith. (Lavyngham, *Litil Tretys*, a1400, 8.21-27)

(27) Ther ben somme that byleue that they haue their destyne after the cours of the sterres, the whyche thyng is false and euyl errour. For *Seint Gregorye sayth that* no good crysten ought to byleue that thei haue ony other destynee but that onely God […] be plesed with. (Caxton, *Doctrinal of Sapience*, 1489, 55.19-23)

Mental verbs in the first and third person occur regularly in expository text sections since they may be used to express first-person stance or to describe the stance of third parties. These textual functions are particularly relevant in religious treatises because the texts commonly contain expositions of the author's faith, i.e. statements describing the author's religious beliefs and personal opinions, which are shared by members of his religious denomination, as in example (28), or because the texts present the beliefs of others, as exemplified in excerpt (29).

(28) *We þinkis þat* […] qwher-euer we be, sytt we stand we, dreyd of god fro our hartis passis nott. (Misyn, *Mending of Life*, 1434, 106.17-20)
(29) And god suffred them by their naturall reason to come to the knowledge of many secrete & inuisible perfeccions of his diuinite or godhed (as saynt Poule sayth). Wherefore syth *they knewe that* ther is one and euer hath ben / that is the first principall causer of all thynges. (Bonde, *Pilgrimage of Perfeccion*, 1526, 4v.21-26)

In argumentative contexts, nominal clauses occur for the following reasons. First-person mental verbs are often used to express the author's opinion, which is contrasted with opinions held by other religious groups, as can be seen in example (30). Moreover, in passages that contain third-person communication and mental verbs, deviant statements or views of members of other religious denominations are commonly introduced as counter-arguments to the main thesis of the text. Thus, they are first contrasted with the views of the author and later refuted in the author's line of argumentation, as in example (31). The two excerpts below both come from authors with Protestant backgrounds, who argue against Catholic beliefs.

(30) Thirdly the holy Ghost annouincheth, Eph.2.2. Coloss.2.13. that all men by nature are dead in sinnes and trespasses: not as the Papists say, weak, sick, or half dead. *Hence I gather, that* man wanteth natural power not to will but freely and franckly to will that which is good. (Perkins, *Reformed Catholicke*, 1597, 22.3-9)
(31) But now, alas and alas again, *the false prophets of Antechrist* which are past all shame, *do openly preach* in pulpets vnto the people of God *that* the catechisme is to be counted heresy: whereby theyr olde blindnesse is brought home agayne. (Ridley, *Pituous Lamentation*, 1566, int2.12-18)

Moreover, the decrease in second-person mental verbs in EModE treatises needs to be commented on. Constructions with second-person mental verbs may be used in the indicative or the imperative mode. In each case, they are used to address the readers directly and to guide them in their thought processes. A decline in these constructions in EModE shows that religious instruction takes a less direct form, with the writer using more subtle means of instruction. Excerpt (32) is taken from an exposition of the Pater Noster; (33) from an exposition of the Creed.

(32) And *þou shalt vndirstonde þat* fyue þingis letten preier of God to be herd. þe first is yuel lijf of þe preyande… (Ermyte, *Pater Noster*, c1425, 6. 40-41)

(33) The second [article of the faith] is […] "I beleue in Ihesu Cryst his oonly Sone, our Lord". *Vnderstonde here that* Ihesu Cryst the Sone is euen wyth the Fader wythout begynnyng and that the Fader doth nothynge wythout the Sone ne the Sone wythout the Fader. (anon., *Quattor Sermones*, 1483, 22.22-25)

Altogether, the analysis of verbal constructions controlling nominal *that*-clauses in religious treatises indicates that some of the communicative functions of nominal clauses make them especially valuable to writers of religious texts. They are practically unavoidable in the exposition of different beliefs and statements, or the discussion of religious matters by contrasting different points of view. The increase in particular verbal constructions (especially third-person communication, but also first-person mental verbs) also suggests an increasingly argumentative style of writing in religious treatises. As discussed above, the diversification of beliefs during the Reformation and thus the greater number of controversial topics addressed in religious treatises clearly triggered the increased need to contrast different opinions in religious texts.

Interestingly, zero *that*-clauses also rise notably in religious treatises in the EModE period.[6] The omission of the subordinator is typical of informal writing, as it results in a less explicit expression of syntactic relations and therefore an uncertain presentation of information (Biber 1988: 196). Therefore, writers of religious treatises appear to adopt a more informal style of writing. *That*-clauses with an ellipted subordinator can be found especially in persuasive texts, which are addressed to a broad, largely unlearned audience and which are therefore written in a simple style. They can additionally be found in polemical texts, which are designed to convince as many readers as possible of particular points of dispute. Baxter's *Call to the Unconverted*, in which he urges the common man to repent and convert to God, is an example of the former type of treatise (see example (34)), and Goodman's *Compassionate Inquiry*, in which he urges dissenters to conform to the ways of the established Church of England (example (35)) constitutes an example of the latter type.

(34) Others will think, Its true that we must Turn from our evil waies, but I am Turned long ago; *I hope Ø* this is not now to do. And thus while *wicked men think Ø* they are not wicked, but are already Converted, we lose all our labour in perswading them to Turn. (Baxter, *Call to the Unconverted*, 1658, 26.31-27.5)

(35) To this therefore I answer in the second place, That *it is certain Ø* all is not to be esteemed Popery, that is held or practised by the church of Rome, and it cannot be our duty to depart further from her than she hath departed from the truth: for then it would be our duty to forsake Christianity it self in detestation of Popery. (Goodman, *Compassionate Inquiry*, 1674, 12.8-15)

The detailed analysis of verbal constructions controlling nominal *that*-clauses in prayers presents a different picture (see table 6). In prayers, first-person communication verbs controlling nominal clauses show consistently high frequencies in ME and in EModE. Additionally, first-person mental verbs have high frequencies, especially in ME, while second-person communication verbs increase markedly in EModE.

Table 6: Distribution of verbal constructions controlling nominal *that*-clauses which function as complements to verbs in religious treatises (absolute numbers and relative frequencies per 10,000 words)

	ME 13,581 words		EModE 61,109 words	
Subordinators with:				
1st person mental verbs	17	12.5	31	5.1
2nd person mental verbs	2	1.5	29	4.7
3rd person mental verbs	1	0.7	9	1.5
1st person communication verbs	23	16.9	90	14.7
2nd person communication verbs	15	11.0	148	24.2
3rd person communication verbs	0	0	5	0.8

Some 95.6 percent of first-person communication verbs in prayers serve as explicit performatives, in which the speech act expressed is made explicit by the performative verb in the matrix clause. Prayers may therefore be considered a highly performative genre (see also Kohnen forthcoming). Since prayers can be defined as addresses to God or a saint which contain requests, thanksgivings, confessions, and praise,[7] it is not surprising that the explicit performatives found in prayers constitute, for the most part, these types of speech acts. The aim of making speech acts explicit is to heighten the clarity of utterances by ruling out any type of ambiguity in what is being expressed, thus heightening the force of

the utterances. The avoidance of ambiguity, as well as increased force, are clearly desirable attributes of communication with God.

Requests to God are the most common type of speech acts made explicit in prayers. As can be seen in examples (36) and (37), the requests centre predominantly on deliverance from evil temptations or sins, and consequently a life in peace, following God's word. These requests apply to all Christians. Requests which pertain to an individual's particular needs, such as deliverance from personal grief, e.g. illnesses or other types of hardship, are rarely expressed. This has to do with the fact that prayers preserved in the written form were largely adapted for use by a larger group of believers.

(36) *We biseche þee*, almiȝti god, *þat* bi þe meritis of þo modir & maide marie, & of alle halewene, we be defendid from alle yuelis. so þat þorouȝ her preieris we moun lyue peisibli in þi worschip, bi crist oure lord. amen! (anon., *Hours Blessed Virgin*, c1400, 14.30-33)

(37) Wherfore of synguler grace and mercy *I beseche thee that* fro this day foreward I go nevere out fro thi doctryne, that thin eendelees goodenesse hath taught me. (anon., *Selected Prayers*, c1425, 98.8-10)

Giving thanks is another type of speech act which is often made explicit in prayers. Of course, people praying do not address God only to make demands. They also wish to thank him for the things he has already granted. Again, gratitude to God typically denotes things with a general application for all Christians: God's divine mercy and grace, which preserved the person praying in times of need or trouble (see examples (38) and (39)).

(38) And *with all my herte I thanke the* most mercyful lorde / for the grete mercyes that thou haste shewed me in the grete daungers that I haue ben in as wel in my soule / as in my body / and *that* thy grace and endles mercy hath euer kept me, spared me and saued me from the hour of my byrthe in to this tyme. (anon., *Various Devout Prayers*, a1500, int14.7-15)

(39) *I thank thee, that* thou hast brought this weeke about with me, helping mee with all things which were needfull for my Body. (Sparke, *Crums of Comfort*, 1628, int15.32-35)

Explicit performatives are also used for the confession of sins (see (40) and (41)). The sins confessed in these prayers also do not constitute particular sins committed by the individual praying (e.g. theft, adultery, murder) but, instead, they often treat the sinful and lowly nature of all of mankind. In example (41), the person praying is additionally praising God by means of an explicit performative.

(40) But first of all, *wee doe here confesse, that* we were conceiued in sin, and brought forth in iniquitie; And by reason of our original corruption, drawne from the loynes of our Parents, we are apt to euerie thing, that is euill. (Sparke, *Crums of Comfort*, 1628, int13.2-5)

(41) O Almighty God, *I acknowledge and confess that* I am less than the least of all thy Mercies, and am unworthy of the least Crum that falls from the Table of thy ordinary Providence. *I praise and bless thy glorious Name, that* thou hast preserved me from the dangers and perils of this night, and continued to me still the opportunities of serving thee. (Hove, *New-Years-Gift*, 1681, 5.18-6.12)

The other type of verbal construction introducing nominal clauses which occurs with high frequencies in prayers and increases noticeably in the EModE period is second-person communication verbs. A closer look at these verbal constructions shows that 93.3 percent constitute imperatives. Imperatives typically function as directives, which are, according to Searle, attempts by the addressor to get the addressee to carry out a particular action (1969: 66; 1976: 11). As mentioned above, one of the main purposes of praying is to ask for assistance from God in various matters. These requests manifest themselves in the form of explicit performatives (see examples (36) and (37) above) or in the form of imperatives, as can be seen in the examples below (examples (42) and (43)). Again, the things asked for have a general application to all Christians, asking God to grant a life according to the doctrine, rather than dealing with more individual concerns.

(42) *Grant* vnto vs *that* neither this day, nor at any time, any euill may take hold of vs. (Sparke, *Crums of Comfort*, 1628, int16.33-36)

(43) *Graunt* to me *that* I may so confesse the my lord god with my mouth that I neuer do contrary thy preceptes in my dedes. (anon., *Deuoute Prayers*, 1535, 13.1-3)

Lastly, table 6 (above) shows that first-person mental verbs have a high frequency in the ME period but decline, however, in EModE. First-person mental verbs controlling nominal clauses are generally used to express first-person stance, i.e. the private thoughts of the person praying. Especially in ME prayers, these consist of admissions of guilt (example (44)) or descriptions of a troubled conscience (example (45)).

(44) Moost mercyfull lorde Ihesu, *I know well that* I haue ofte synned dedely, both by wyll & dede, wherby ye myght by rightful Jugement haue condempned me forthwyth into helle without ende. (anon., *Prayers Love*, a1500, 105.32-106.2)

(45) Lorde, I will knowlage vnto the, al mine vnrightuousnesse, and I wyll confesse to the all the vnstablenesse of my herte. Oftentymes a veraie litle thyng troubleth me sore, and maketh me dull and slow to serue the. And sometyme I purpose to stande strongly, but whan a litle truble cometh, it is to me great anguish & grief, and of litle thyng riseth a greuous temptacion to me. Yea *when I thinke* my self to be sure and strong, and *that* (as it semeth) I haue ye vpper hand: sodenly I feele my self ready to fall with a little blast of temptacion. (Parr, *Prayers or Meditacions*, 1545, Av.13-20)

In EModE prayers, these verbal constructions are not as common, when they do occur, they are used as an expression of the most fundamental religious belief of the person praying. With the emergent variety of denominations, the person praying must have felt a greater need of expressing his own beliefs as a means of setting himself apart from members of other denominations. In excerpt (46), Edward Dering, a Puritan Non-Conformist, expresses his view of God as being revengeful and merciless, a ruler who will justly punish the wicked. In contrast, the following excerpt (example (47)) is taken from a prayer collection by Nicholas Themylthorp, in which God is portrayed as forgiving and merciful. Considering the number of prayers which he adopted from the official *Book of Common Prayer*, Themylthorp is assumed to have been an Anglican Conformist (Green 2000: 252).

(46) And againe (O Lord) *I see* thy heauie wrath, vengeance, and iudgement against sinne to be intollerable, *that* euen the least wicked thought and most secrete cogitation of my hart, procureth thy heauy wrath, and euerlasting curse, the torments of hell, and euerlasting fire, euen although I had but once in all my life broke any of thy commaundements, so much as once in thought. *I know* (O Lord GOD) *that* thou are true and iust, and canst not abide sin and wickedness, but wilt iustlie punish euery sinne, euen with the selfe-same torments of hell, which thy iustice hath appointed, euen for euery sinne. (Dering, *Godly Private Prayers*, 1597, ll. 2.18-27)

(47) And yet, because I feele in my selfe so manie faults and imperfections, such readinesse to evill, and slacknesse to doe good, I quake, and tremble, and feare of thy wrath, and sharpe judgement: but for that *I know* thou commandest mee by prayer to craue of thee all things necessaie for soule and bodie; and hast promised, graciouslie to heare my lamentable sute, and merciefullie to grant mee my needfull requests and petitions. (Themylthorp, *Posie*, 1636, 10.5-11.4)

In all, the analysis of verbal constructions controlling nominal *that*-clauses in prayers shows that communicative functions which differ from the ones they serve in religious treatises make these constructions good choices for the expression of communicative purposes shared by people praying. Communicative functions include the making explicit of speech acts such as uttering requests, giving thanks and confessing sins to ensure clarity and to avoid ambiguity when addressing God. Additionally, a minor function is the expression of personal stance, usually for the admission of guilt or, in EModE prayers, to express personal beliefs.

4. Conclusion

This pilot study of the connective profiles in religious treatises and prayers has revealed complex constellations of subordinating conjunctions in the two genres. Both similarities and differences could be identified. The increase in subordinators expressing conditional, concessive and causal relations, which coincides with a decline in temporal subordination, is the most remarkable similarity found in the two genres. This development has already been described by Kortmann (1997a and b) as a diachronic universal, which occurs in a variety of European languages and which may be linked to the rise of written standard languages. The present analysis makes clear that a closer look at language use differentiated by genre offers additional insights. Changes in the distribution of subordinators may follow universal paths, but, to a certain extent, they are also genre-specific. Shifts in genre conventions, which can be triggered by changes in the particular socio-cultural background, also influence the choices writers make in their use of clausal connectives. Particularly in religious treatises, the increase in conditional and causal clauses during the EModE period is seen to have been caused by changes in genre conventions, as writers of religious treatises adopted a more elaborate and persuasive style of writing when the need for discussing controversial topics in religious writing grew as a result of the political and socio-cultural developments of the time.

Differences in the connective profiles clearly underline the genre-specific distribution. The divergent development of clauses of purpose and result in the two genres can be tied to various functional motivations, which lose or gain in importance due to genre-specific changes in styles of writing. Whereas the high frequencies of subordinators in clauses of purpose and result in prayers, for instance, can be tied to a heightened desire to express the purposes of particular requests to a transcendental being, purposive and resultative subordinators first appear in early religious treatises as a means of presenting aspects of the doctrine and later in argumentative sections of the text for the discussion of more elaborate religious matters.

Further similarities in the distribution of clausal connectives in the two genres, such as the similar ratio of nominal and adverbial subordinators, do not necessarily contradict a genre-specific distribution. The uniform distribution of nominal subordinators has been shown to only superficially conceal differences in usage. A detailed analysis of the verbal constructions governing nominal clauses revealed that the motivations affecting the use of nominal subordinators differ greatly in the two genres. While they are used for the exposition or discussion of religious beliefs in religious treatises, they are needed in prayers to make speech acts, such as requesting, thanking, confessing, and praising, explicit.

Although the present study can only offer preliminary results, which need to be confirmed by more comprehensive analyses, the contrastive analysis of the connective profiles found in religious treatises and prayers has already offered some intriguing results, not only with regard to genre-specific functional motivations for differences and similarities in the connective profiles, but also

with regard to genre-specific changes in the distribution of particular types of subordinating conjunctions. This suggests that the influence of genre conventions should be accounted for in the distribution of subordinating conjunctions.

Notes

1 In a preliminary count, only ten percent of non-finite clauses were introduced by an explicit subordinator in EModE religious treatises.

2 Note that the category 'other clauses' has not been included in tables 3 and 4 since the frequencies are relatively low and will therefore not be discussed further (religious treatises: ME 34/0.5 and EModE 25/0.4; prayers: ME 1/0.1 and EModE 4/0.1).

3 I would like to thank Tanja Rütten-Stanelle for sharing her valuable thoughts on the history of the genre religious treatises.

4 The distinction corresponds roughly to Quirk et al.'s private and public types of factual and suasive verbs (Quirk et al. 1985: 1180-1182).

5 See the definition of the term 'treatise, n.' given in the *Oxford English Dictionary*: 'book or writing which treats of some particular subject' and which commonly contains a 'formal or methodological discussion or exposition of the principles of the subject'.

6 Increase of zero *that*-clauses from 14/2.2 in ME to 72/11.5 in EModE.

7 The definition of 'prayer' given in the *Oxford English Dictionary* is slightly more restricted, defining a prayer as 'a solemn request to God, a god, or other object of worship, a supplication or thanksgiving addressed to God or a god'.

References

Barratt, A. (1986), 'Works of religious instruction', in: A.S.G. Edwards (ed.) *Middle English Prose. A Critical Guide to Major Authors and Genres.* New Brunswick: Rutgers University Press. 413-428.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, D. (1988), *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Claridge, C. (2007), 'Conditionals in Early Modern English texts', in: U. Lenker and A. Meurman-Solin (eds.) *Connectives in the History of English.* Amsterdam: John Benjamins. 229-255.

Claridge, C. and T. Walker (2001), 'Causal clauses in written and speech-related genres in Early Modern English', *ICAME Journal*, 25: 31-63.

Green, I. (2000), *Print and Protestantism in Early Modern England*. Oxford: Oxford University Press.

Jucker, A.H. (1991), 'Between hypotaxis and parataxis. Clauses of reason in *Ancrene Wisse*', in: D. Kastovsky (ed.) *Historical English Syntax*. Berlin and New York: Mouton de Gruyter. 203-220.

King, J. (2000), 'Religious writing', in: A. Kinney (ed.) *The Cambridge Companion to English Literature 1500-1600.* Cambridge: Cambridge University Press. 104-131.

Kohnen, T. (2007), 'Connective profiles in the history of English texts: aspects of orality and literacy', in: U. Lenker and A. Meurman-Solin (eds.) *Connectives in the History of English*. Amsterdam: John Benjamins. 289-309.

Kohnen, T. (forthcoming), 'Prayers in the history of English'.

Kortmann, B. (1997a), *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages*. Empirical Approaches to Language Typology 18. Berlin: Mouton de Gruyter.

Kortmann, B. (1997b), 'Typology and language change', in: U. Böker and H. Sauer (eds.) *Anglistentag 1996. Dresden*. Trier: Wissenschaftlicher Verlag Trier. 109-124.

Lenker, U. (2007), 'Forwhi 'because': shifting deictics in the history of English causal connection', in: U. Lenker and A. Meurman-Solin (eds.) *Connectives in the History of English.* Amsterdam: John Benjamins. 193-229.

Molencki, R. (2007), 'The evolution of *since* in medieval English', in: U. Lenker and A. Meurman-Solin (eds.) *Connectives in the History of English.* Amsterdam: John Benjamins. 97-115.

Oxford English Dictionary. (1989) 2nd ed. OED online. Oxford: Oxford University Press (http://dictionary.oed.com/ cgi/entry/50256967).

Prince, E. (1981), 'Towards a taxonomy of given-new information', in: P. Cole (ed.) *Radical Pragmatics.* New York: Academic Press. 223-255.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A Comprehensive Grammar of the English Language.* London: Longman.

Rissanen, M. (1989), 'The conjunction *FOR* in Early Modern English', *NOWELE*, 14: 3-18.

Rissanen, M. (1996), 'Genres, texts and corpora in the study of medieval English', *Anglistentag 1995 Greifswald: Proceedings.* Tübingen: Niemeyer. 229-242.

Rissanen, M. (1998), 'Towards an integrated view of the development of English: notes on causal linking', in: J. Fisiak and M. Krygier (eds.) *Advances in English Historical Linguistics.* Berlin and New York: Mouton de Gruyter. 389-406.

Rissanen, M. (1999), 'Syntax', in: R. Lass (ed.) *The Cambridge History of the English Language. Volume III: 1476-1776.* Cambridge: Cambridge University Press. 187-331.

Rissanen, M. (2007), 'From oþ to till: early loss of an adverbial subordinator', in: U. Lenker and A. Meurman-Solin (eds.) *Connectives in the History of English.* Amsterdam: John Benjamins. 61-77.

Searle, J.R. (1969), *Speech Acts*. Cambridge: Cambridge University Press.

Searle, J.R. (1976), 'A classification of illocutionary acts', *Language in Society*, 5: 1-24.

Sorva, E. (2007), 'Grammaticalization and syntactic polifunctionality: the case of *albeit*', in: U. Lenker and A. Meurman-Solin (eds.) *Connectives in the History of English.* Amsterdam: John Benjamins. 115-145.

Appendix

The following texts are taken from the *Corpus of English Religious Prose* (COERP), which is currently being compiled at the University of Cologne.

Religious Treatises: anon., *Speculum Christiani* (c1300); Dan Michel, *Ayenbite of Inwit* (1340); Richard Rolle, *Forme of Liuing* (a1349); anon., *Mirror of St. Edmund* (1350); John Wyclif, *Of Clerkis Possessioneris* (?c1370-1400); Walter Hilton, *Eight Chapters on Perfection* (?a1396); anon., *Cloud of Unknowing* (?a1400); Richard Lavyngham, *Litil Tretys on Seven Deadly Sins* (a1400); Nicholas Love, *Mirror of the Blessed Life of Jesus Christ* (a1410); anon., *Þe Lanterne of Li3t* (a1415); Richard Ermyte, *Treatise on Pater Noster* (c1425); Richard Mysin, *Mending of Life* (1434); anon., *Craft of Dying* (c1450); anon., *Quattor Sermones* (1483); William Caxton, *Doctrinal of Sapience* (1489); anon., *Tretyse of Loue* (1493); William Bonde, *Pilgrimage of Perfeccion* (1526); William Tyndale, *Obedience of a Christen Man* (1528); anon., *Christe and his Teachings Compared with Pope and his Doings* (1534); William Baldwin, *Treatise of Moral Philosophiae* (1547); Nicholas Ridley, *Pituous Lamentation* (1566); John Northbrooke, *The Poor Mans Garden* (1571); William Perkins, *A Reformed Catholicke* (1597); Robert Linaker, *A Comfortable Treatise* (1595); Thomas Tymme, *Silver Watch-bell* (1605); Humphrey Lynde, *Via Tuta: The Safe Way* (1628), Nicholas Ferrar, *Little Gidding* (1631); Ephraim Pagitt, *Heresiography* (1645); Richard Baxter, *Call to the Unconverted* (1658); John Goodman, *Serious and Compassionate Inquiry* (1674); John Tillotson, *Discourse against Transubstantiation* (1684); John Rawlet, *The Christian Monitor* (1686).

Prayers: anon., *Two Middle English Prayers* (a1400); Thomas Aquinas, *Oracio Thome de Alquin* (a1400); anon., *The Hours of the Blessed Virgin* (c1400); anon., *Pater Noster & Ave Maria* (c1400); anon., *Selected Prayers* (c1425); anon.,

Fifteen Oes (1491); anon., *Various Devout Prayers* (a1500); anon., *Prayers in Treatise of Love* (a1500); anon., *Diverse EModE prayers* (c1510); anon., *Mystic Rosary* (1533); anon., *Deuoute Prayers in Englysshe* (1535); Catherine Parr, *Prayers or Meditacions* (1545); Thomas Becon, *Pomaunder of Prayer* (1561); anon., *Fourme of Prayer* (1570); Thomas Bentley, *The Fifth Lampe of Virginitie* (1582); Edward Dering, *Godly Private Prayers* (1597); Thomas Sorocold, *Supplications of Saints* (1612); Michael Sparke, *Crums of Comfort* (1628); Nicholas Themylthorp, *Posie of Godly Prayers* (1636); anon., *Three and Thirty Godly and Devout Prayers* (1640); Jeremy Taylor, *The Golden Grove* (1654); Church of England, *The Primer* (1658); Ann Morton, *Daily Exercise* (1666); Frederick van Hove, *A New-Years-Gift* (1681).

A contrastive look at English and Dutch (negative) imperatives

Daniël Van Olmen

University of Antwerp

Abstract

Somewhat surprisingly, the imperatives of the Germanic neighbor-languages of English and Dutch have not yet been compared in a systematic, corpus-based way. This paper is a first step towards such a contrastive study. Firstly, it looks at the frequencies of imperative subtypes in the spoken part of the International Corpus of English – Great Britain *and in a comparably compiled Northern Dutch corpus out of the* Spoken Dutch Corpus*. These quantitative results raise questions about the use of imperative discourse markers in both languages, about the grammaticalization of hortatives and about alternative linguistic means of expression. A second section focuses on negative imperatives in English and Dutch. It provides a pragmatic analysis of prohibitives from the perspective of speech act theory and examines their translations in a two-way parallel corpus of plays. The English and Dutch negative imperatives are found to have roughly the same illocutionary profile. The parallel corpus data reveals a difference in correlation between prohibitives in both languages. It is argued that this distinction is part of the explanation for the frequency facts.*

1. A first look

In line with van der Auwera (2006: 565), the imperative is considered here as a construction of grammar which typically presents a state of affairs as desirable by the speaker and calls on the hearer(s) to actualize it. For English and Dutch, this characterization entails that examples such as (1) and (2) cannot be regarded as genuine imperatives.

(1) a. I want you to leave!
    b. Ik wil dat je vertrekt!
(2) a. I'd leave, if I were you.
    b. Ik zou vertrekken, als ik jou was.

The sentences in (1) satisfy the conditions of a desirable state of affairs and of an appeal to the addressee, but they use lexical means, more precisely the volitional verbs *want* and *willen* 'want', rather than grammatical ones. In (2), the call on the addressee is not even part of the sentence meaning. It is the possible outcome of a conversational implicature.

What the grammatical category of the imperative does involve in English is shown in (3): a tenseless verb form, *do*-support for emphasis and negation as in (3b) and (3d), a non-compulsory subject as in (3c) and *let* for hortatives such as (3e).[1]

(3) a. Leave!
b. Do leave!
c. You leave!
d. Don't leave!
e. Let's leave!

The formal characteristics of the Dutch imperative are illustrated in (4). The verb usually appears in sentence-initial position. Its form can be different from that in (4a) but only if the optional subject is overtly expressed as in (4b) (particles like *maar* 'but, feel free to' are often necessary for an imperative with a subject to be acceptable in Dutch). Negation requires the mere addition of *niet* 'not' as in (4c) or of some other negative element. And hortatives make use of the verb *laten*, the Dutch cognate of *let*, as in (4d).[2]

(4) a. Vertrek!
leave.IMP
b. Vertrekken jullie maar!
leave.IMP.PL you.PL PRT
c. Vertrek niet!
leave.IMP not
d. Laten we vertrekken!
let.IMP.PL we leave.INF

In keeping with the above definition of the imperative, the sets of features in (3) and (4) identify a part of grammar which expresses a (negated and/or joint) state or event as wanted by the speaker and which calls on the hearer(s) to carry it out (the speaker may be one of the addressees or even the only one).

Perhaps surprisingly, contrastive studies on the imperative in English and in Dutch are few and far between (Geukens 1986; van der Auwera and Taeymans 2004). Corpus linguistic research on the topic is simply non-existent. This paper is a first step towards a thorough, corpus-based investigation into the similarities and the dissimilarities between both imperatives.

2. A quantitative look

The main question in this section is whether or not English and Dutch differ with respect to the frequencies of the various imperative forms. Further issues are: to what extent do they diverge and how can the differences be explained?

2.1 Comparable corpora

For a sensible quantitative comparison of English and Dutch, we inevitably have to resort to comparable corpora or, to be precise, "corpora of comparable texts in different languages" (Johansson 1998: 5). The English part of the present corpus is the entire spoken component of the *International Corpus of English – Great Britain* (ICE-GB). The imperatives in this corpus have already been counted and

analyzed by De Clerck (2006). The Dutch part is made up of material from the *Spoken Dutch Corpus* (CGN) and it follows the design of the ICE-GB. A couple of remarks are in order, however.

- Only the imperatives in the parsed files of the nine million-word CGN can be searched for electronically. Unfortunately, these syntactically annotated files account for only ten percent of the data.
- The Flemish or Southern Dutch component, one third of the whole corpus, cannot be used in view of the parallel corpus material (see section 3.2).
- The CGN does not possess many monologues. Especially scripted ones are in short supply. Since we want to have the same proportion of monologue and dialogue as in the ICE-GB, this means that the total number of Dutch words must be (much) lower in the corpus.
- Even so, it is unavoidable that the scripted monologues are under-represented in the Dutch corpus. The unscripted monologues have to make up for it.

Table 1 below provides the approximate number of words and the percentage of the total number of words for each subcorpus in the English and the Dutch parts of the corpus (see Nelson, Wallis and Aarts 2002 for the text types in the ICE-GB).

Table 1: A comparable corpus of spoken English and Dutch

		English		Dutch	
Dialogue	Private	200,000	33%	100,000	33%
	Public	160,000	27%	80,000	27%
	Total	360,000	60%	180,000	60%
Monologue	Unscripted	140,000	23%	90,000	30%
	Scripted	100,000	17%	30,000	10%
	Total	240,000	40%	120,000	40%
Total		600,000	100%	300,000	100%

2.2 Findings

The frequencies of the imperative subtypes and their distribution over the various subcorpora are given in figures 1 and 2. They sum up De Clerck's (2006: 202ff.) British English results and our Northern Dutch results for, respectively, dialogues and monologues. What is important to note, however, is that they only distinguish between positive imperatives (+), hortatives (*let*) and prohibitives (-). Emphatic positive imperatives and hortatives are subsumed under the general labels because there is no distinct equivalent of this type of *do*-support in Dutch and because the phenomenon is uncommon (about one percent of all the imperatives in the spoken ICE-GB). Low frequency is also the reason why positive and negative imperatives with subjects (e.g. two percent of the imperatives in Dutch) do not receive a separate mention in the figures.[3]

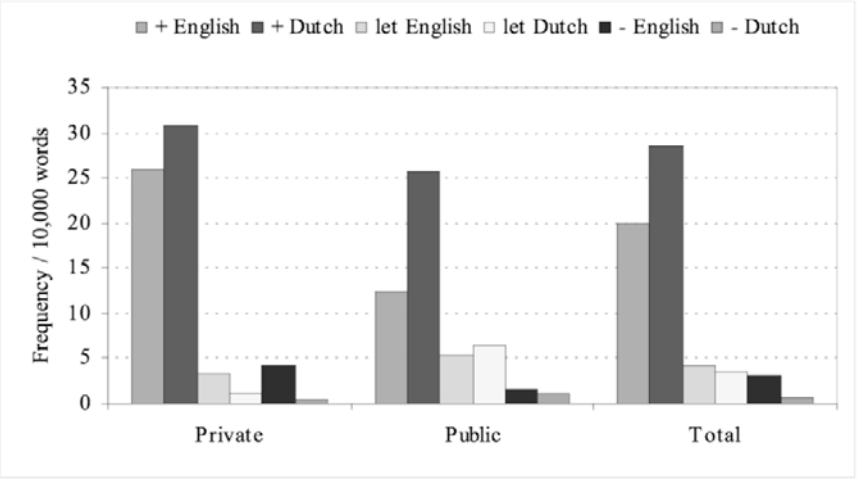

Figure 1: The imperative subtypes in English and Dutch dialogues

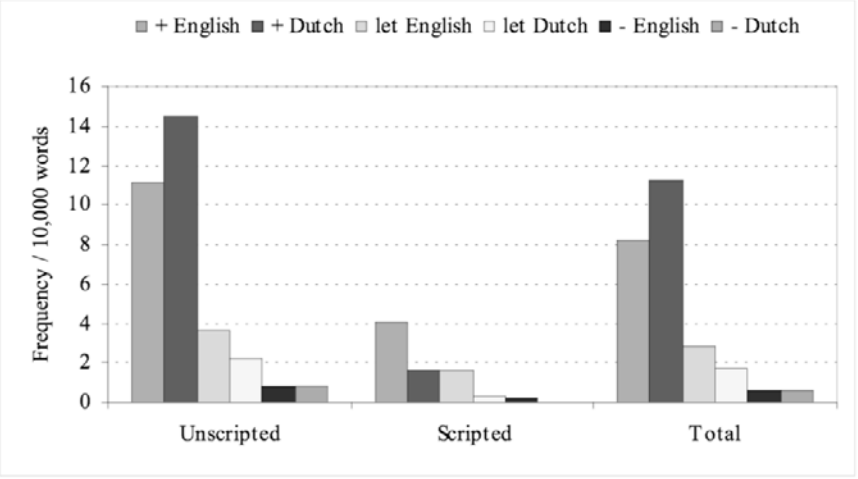

Figure 2: The imperative subtypes in English and Dutch monologues

With regard to the text types, it is hardly surprising that the dialogue subcorpus contains twice as many imperatives as the monologue subcorpus (note the different scales in the figures). Obviously, imperatives can be used as references, recommendations, instructions and expository directives as shown, in that order, in (5) to (8). In these examples, the imperative does not actually seem to need an (active) addressee. But on the whole, it can be assumed that imperatives, by their

very nature, involve interaction and that they are thus more likely to occur in dia- than in monologual contexts.

(5) Zie de motie Remkes. (CGN: fn000177.90)
see.IMP the bill Remkes
'See the Remkes bill.'
(6) Never mind the quality feel the width. (ICE-GB: S1A027.281)
(7) Well a couple of minutes swishing it to and fro to and fro and then don't wash out with anything afterwards. (ICE-GB: S1A087.288)
(8) Neem 't woord leefbaar. (CGN: fn007125.8)
take.IMP the word bearable
'Take the word *bearable*.'

The difference in frequency between the two kinds of monologue is also in line with expectations. The unscripted genres of, among others, demonstrations and live commentaries are still moderately interactive. But the scripted monologue of, for instance, broadcast news is simply not compatible with manifestations of the speaker's desire and calls on the hearer(s). Interestingly, this distinction is much more substantial in Dutch than in English. The overall frequency of imperatives in unscripted monologue is twice that in scripted monologue in English but nine times greater in Dutch. The reason(s) for this dissimilarity can only be brought to light by an in-depth analysis of the imperative's pragmatic functions and also of the genre conventions in the two languages.

The comparison of the imperative subtypes reveals similar hierarchies in English and Dutch. Positive imperatives are much more common than hortatives, which, in turn, are more frequent than negative imperatives. In the subcorpora, there are just two exceptions to this 'pecking order'. Firstly, the positive imperatives in scripted monologue do not overshadow the hortatives as much as they do in the other text types. Secondly, the English hortatives in private dialogue are slightly less frequent than the negative imperatives. Again, a careful genre and speech act analysis is necessary to account for these observations.

The fact that English and Dutch have the same hierarchy of subtypes does not, of course, mean that they are exactly alike. In general, there appear to be more imperatives in Dutch than in English. But the difference can in fact be traced back to a single subtype; the positive imperative, by far the most common subtype in either language, is one and a half times more frequent in Dutch than in English. The higher number in Dutch is the result of the relative proliferation of so-called imperative discourse markers in the language. The distinction between *look* and *kijk* 'look' is a case in point (Van Olmen 2008). In the spoken section of the ICE-GB, there are 32 attestations of *look* in contexts like (9). The Dutch corpus, which is half the size of the English one, contains 114 examples of *kijk* with a discourse use as in (10). Tellingly, the metapragmatic comment *zeg maar* 'say' in (11) is also (nearly four times) more frequent than its English counterpart in (12); in both examples the imperative discourse marker functions as an approximator.

(9) And I said look you're not to cry. (ICE-GB: S1A094.111)
(10) Nou uh kijk competitie doe ik niet meer. (CGN: fn000868.37)
now uh look competition do.PRES.1SG I not anymore
'Now uh look, I no longer play competition.'
(11) … dat ik eerst zeg maar twee uurtjes thuis ga leren. (CGN: fn000573.197)
that I first say.IMP PRT two hours home go.PRES.1SG study.INF
'… that I'll study at home first for, say, two hours.'
(12) OK we can guarantee say a thousand barrels of oil per day over this particular route. (ICE-GB: S1B005.48)

In contrast, the English hortatives and negative imperatives exhibit higher rates of occurrence than the Dutch ones. The hortative construction, to start with, is thirty percent more frequent in English than in Dutch. In light of the fairly high degree of grammaticalization of *let's* (De Clerck 2006: 220-238; Hopper and Traugott 1993: 10-14), as evidenced by (13), and of the compositionality of the structure in Dutch, one may ask whether the English hortative has very (inter)subjective uses that its Dutch counterpart does not have. This question must remain unanswered here. What is also a topic for further research is the role of (14) in the difference in frequency between the English and the Dutch hortatives. The combination of one or both of the particles *eens* 'once, sometime' and *even* 'briefly, just' with the infinitive *kijken* 'look' is extremely widespread in the Dutch material. It seems to function in more or less the same way as the formulaic expressions *let me see* and *let's see* (as a hesitation marker, for instance).

(13) The UN Motto: "*Lets you and him start conserving fossil fuel!*" (internet example from De Clerck 2006: 231)
(14) Eens/even kijken.
PRT look.INF
'Let me/'s see.'

The frequency of the English negative imperative, then, is over three times that of the Dutch one. This huge difference raises the following questions: is the English negative imperative used for more pragmatic purposes than the Dutch one and/or does Dutch use other linguistic means to express the functions of the prohibitive and, if so, which ones? These issues are addressed in the next section. Note, though, that the distinction between English and Dutch is mainly due to the private dialogue subcorpus. In the (un)scripted monologue and public dialogue components, the frequency of the Dutch negative imperative is identical to, or just slightly lower than, that of its English counterpart. But in the private dialogue component, the frequency of the Dutch negative imperative is not even one-tenth that of the English one.

3. A negative look

3.1 Parallel corpora

The use of parallel or translation corpora in this paper is prompted by the question about alternative linguistic means of expression to the negative imperative. This type of corpus may provide some answers since, as Aijmer and Altenberg (1996: 13) claim, "translations make it possible to investigate how the same content is expressed in two languages" (see Mauranen 2002: 167 for a critical assessment). But there are also a number of disadvantages to the approach.

- There is still no parallel corpus English ↔ Dutch that is (easily) available. The corpus here is thus a set of texts collected specifically for the present investigation.
- The range of genres that are well translated is very limited. What is more, most translated texts that are acceptable are not helpful for a study of the prohibitive anyway. Legal documents, for one, can be assumed not to have enough imperatives. In view of the comparable corpora and the findings in the preceding section, the texts must resemble (private conversational) spoken language. Dramatic texts with 'everyday topics' probably come closest to meeting the criteria.
- The data is neither tagged nor parsed, which means that the negative imperatives cannot be searched for electronically.
- For all the above reasons, it cannot be avoided that the parallel corpora are rather small. Luckily, the concentration of negative imperatives appears to be very high in plays.

The English → Dutch corpus is made up of five plays by different British authors and contains about 97,000 words. Each play is translated by a different translator and like nearly all texts, for cultural and economic reasons, into Northern Dutch.[4] The Dutch → English corpus is compiled in an analogous manner and has a word count of just over 70,000.[5]

3.2 Pragmatic analysis

The central issue here is whether the negative imperative functions differently, i.e. occurs in other speech acts, in English and Dutch. The analysis that can offer an answer to that question calls for a detailed taxonomy of illocutionary goals. Unfortunately, but understandably (any categorization of the pragmatics of a linguistic element is a simplification of real life), there is no agreement on the functions that should (not) be included in such a classification. In this paper, we make use of De Clerck's (2006) refinement of De Rycker's (1990) taxonomy of speech acts because it allows for a broader characterization of imperatives. They are not automatically simple directives. In the following table, the major classes are listed and defined. They are illustrated with a few more specific, self-evident speech acts (chiefly in English, for the sake of convenience).

Table 2: De Clerck's (2006: 148-149) taxonomy of pragmatic functions[6]

Major category	Definition	Examples
Wilful directive	Strong appeal to the hearer, often in hierarchical contexts, to do what the speaker wants and what is usually to the benefit of the latter	Command in (15) and plea in (16)
Non-wilful directive	Less strong appeal to the hearer, in hierarchy-neutral contexts, to do what the speaker thinks is to the benefit of the former	Piece of advice in (17) and warning in (18)
Commissive directive	Weak to strong commitment of the speaker to do something which is often to the benefit of the speaker and the hearer and which usually also involves some action by the hearer	Conditional threat in (19) and conditional promise in (20)
Expressive directive	Appeal to the hearer in which the speaker primarily expresses his or her attitude towards the hearer	Support in (21) and challenge in (22)
Mixed expressive	Some kind of appeal through which the speaker hopes to bring about a state of affairs that the hearer does not control and that shows the speaker's attitude towards the hearer	Wish in (23)

(15) Don't pull Mason onto the punches, says referee Larry O'Connell. (ICE-GB: S2A009.34-35)
(16) Don't make me suffer this. (Bond 1982: 16)
(17) So uh I said to her don't leave it until the two weeks before the performance. (ICE-GB: S1A083.108)
(18) But don't underestimate the problems. (ICE-GB: S2A023.77)
(19) Don't you bloody dare. (Stoppard 1982: 53)
(20) Buy nothing and you may still fly back. (Brenton and Hare 1985: 33)
(21) No don't worry about it really. (ICE-GB: S1A091.144)
(22) Oh, don't be ridiculous. (Ayckbourn 1979: 157)
(23) a. Ga alsjeblieft niet dood. (Herzberg 1989: 57)
b. Please don't die. (Herzberg 1995: 82)

One unintended advantage of the parallel corpora is that the limited data in our comparable corpus of Dutch (just twenty negative imperatives in total) can be complemented by the data in the plays (seventy more negative imperatives). Figure 3 below gives the percentages of the major illocutionary classes for the comparable corpus data (Comp), for the parallel corpus data (Para) and for all

negative imperatives together (Total). The proportions in the first column are based on the analysis in De Clerck (2006: 298ff.).

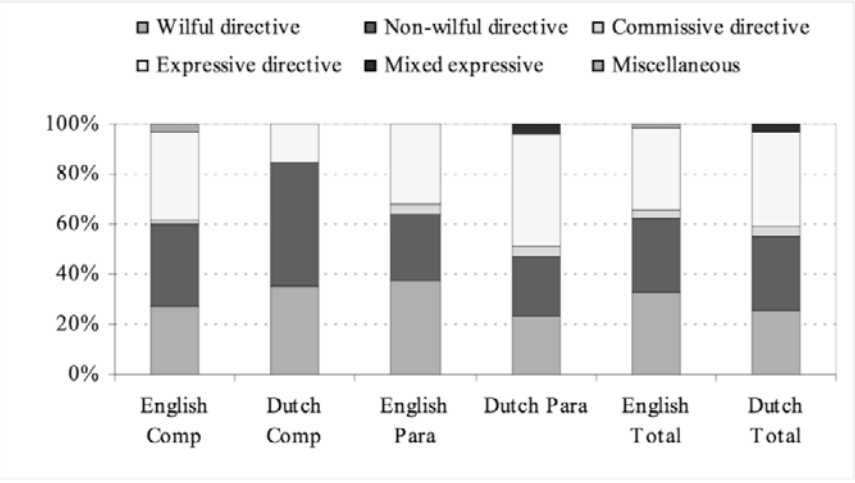

Figure 3: The pragmatic functions of English and Dutch negative imperatives

The final two columns show that, at least at the level of the major classes, English and Dutch negative imperatives are not very different. The expressive and (non-) wilful directives are the dominant illocutionary categories in both languages. The commissive directives and the mixed expressives must be regarded as secondary classes. In other words, there does not seem to be a fundamental distinction in the functional potential of the English and Dutch prohibitives, which might account for the frequency facts. Then again, the Dutch negative imperative is hardly used as an expressive directive (as a support, to be precise) in the comparable corpus, while this class (and the supportive act in particular) is the most frequent one for the English negative imperative. The numbers (for Dutch) are obviously too low to make claims but we can ask: do speakers of Dutch perhaps prefer other linguistic means to the prohibitive for expressive directives?

3.3 Translations

Figure 4 divides the translations of the 126 English negative imperatives and the 70 Dutch ones into two categories: those that take the form of the negative imperative and those that do not. The difference between English and Dutch is clear. Three quarters of the Dutch negative imperatives, but just half of the English ones, are translated into the other language as negative imperatives. The correlation between both prohibitives can thus be said to be stronger from Dutch into English than from English into Dutch. Or, in other words, the parallel corpus findings suggest that speakers of Dutch tend to use alternative means of

expression to the negative imperative to a greater extent than speakers of English (this formulation assumes, and rightly so, that, in principle, all source prohibitives in the data can be rendered in the target language as prohibitives). This result may account for the numbers in figures 1 and 2 above.

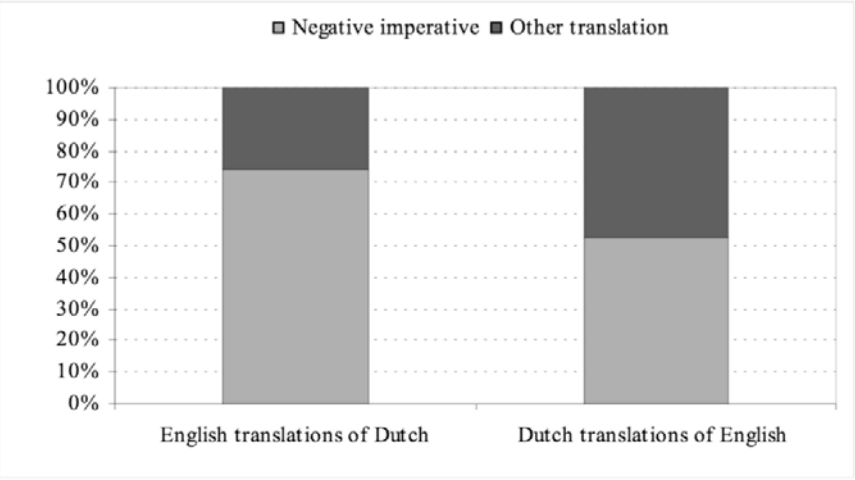

Figure 4: Correlation of negative imperatives in English and Dutch

Moreover, the variety of 'other' translations is much wider for the English source negative imperatives than for the Dutch ones. The translations of the latter include six declarative apologies such as (24), four imperative apologies such as (25), seven other positive imperatives such as (26) and the free translation in (27).

(24) a. Neem mij niet kwalijk. (De Boer 1991: 16)
take.IMP me not detrimental
b. I'm sorry. (De Boer 1997: 31)
(25) a. Neem me niet kwalijk. (Lemmens 1974: 25)
b. Excuse me please. (Lemmens 1978: 25)
(26) a. Zeur toch niet man! (Lemmens 1974: 44)
nag.IMP PRT not guy
b. Oh, shut up, will you. (Lemmens 1978: 85)
(27) a. Vertel mij niks over ze. (Lemmens 1974: 29)
tell.IMP me nothing about them
b. I know them inside out. (Lemmens 1978: 53)

Interestingly, Dutch appears to have a negative imperative as one of its idiomatic ways of apologizing. English, on the other hand, does not. A negative imperative like *don't hold it against me* may sometimes function as an apology but the conventional strategies are exemplified in (24b) and (25b). What is important to note is that the Dutch idiom is actually very formal and rare. Unlike simple *sorry* and

the imperative *excuseer*, the prohibitive in (24a) and (25a) does not even occur in the CGN.

The translations of the negative imperatives in the British plays include a large number of positive imperatives too, such as (28), but also eleven declarative sentences with a modal and a second person subject such as (29), ten independent noun phrases such as (30), ten infinitives such as (31) and miscellaneous forms such as (32) to (34).

(28) a. Don't get kicked by the horse. (Stoppard 1982: 45)
b. Kijk uit dat je geen trap krijgt van een paard. (Stoppard 1984: 55)
look.IMP out that you no kick get.PRS.2SG of a horse

(29) a. Don't thank me. (Pinter 1982: 75)
b. Moet je mij niet voor bedanken. (Pinter 1984: 17)
must.PRS.2SG you me not for thank.INF

(30) a. Don't worry. (Ayckbourn 1979: 192)
b. Geen zorgen. (Ayckbourn 1980: 85)
no worries.

(31) a. And don't say you're just cruising about. (Pinter 1982: 47)
b. En nou niet zeggen dat je zomaar wat rondtourt. (Pinter 1984: 71)
and now not say.INF that you just a.bit cruise.PRS.2SG

(32) a. Think nothing of it. (Brenton and Hare 1983: 28)
b. O, dat was niet. (Brenton and Hare 1985: 30)
oh that be.PST.3SG nothing

(33) a. Don't mess about. (Stoppard 1982: 17)
b. Even serieus. (Stoppard 1984: 26)
briefly serious

(34) a. Don't you bloody dare. (Stoppard 1982: 53)
b. Als je 't lef hebt. (Stoppard 1984: 63)
if you the guts have.PRS.2SG

Two comments are in order here. Firstly, it is remarkable that less than one-third of the 23 supportive acts in English is translated into Dutch as a negative imperative. Even *don't worry* is not translated as a negative imperative in the majority of cases, while it has two straightforward informal prohibitive Dutch counterparts as (35) shows. Possibly, this potential tendency and the relative infrequency of the expressive directive in the corpus of spoken Dutch (see figure 3) partly explain the difference in frequency between English and Dutch and, in particular, the two private dialogue corpora. Incidentally, three quarters of the English prohibitive supports occur in the private dialogue component (De Clerck 2006: 309). They represent more than twenty-five percent of all negative imperatives in that type of text.

(35) a. Don't worry. (Ayckbourn 1979: 190; Stoppard 1982: 84)
b. Maak je maar geen zorgen. (Ayckbourn 1980: 81)
make.IMP yourself PRT no worries
c. Trek 't je niet aan. (Stoppard 1984: 95)
be.concerned.with.IMP it yourself not PRT

Secondly, it should be noted that most of the 'other' translations into Dutch are in fact also acceptable in English. Only the English infinitive cannot function as an alternative to the negative imperative.[7]

4. Summary and outlook

In this comparison of the English and Dutch imperatives, the following observations are made. First, the positive imperative is more frequent in Dutch than in English. The reason for this difference is the proliferation of imperative discourse markers such as *kijk* in Dutch. Second, the hortative is more frequent in English than in Dutch. This fact is hypothesized to be interesting from a grammaticalization/(inter)subjectification point of view and in light of Dutch *eens/even kijken*. Third, the negative imperative is much more frequent in English than in Dutch, more specifically in private dialogue. Fourth, and probably relatedly, the correlation between the English and Dutch negative imperative is stronger from Dutch into English than from English into Dutch. The translations into Dutch, and thus possibly the alternative means of expression in Dutch, are not only more numerous but also more varied than the translations into English. Fifth, and finally, the functional potentials of both negative imperatives are very similar.

In further research, we plan to examine the frequencies in the comparable corpora of the alternatives attested in parallel corpora. We also intend to set up acceptability judgment and elicitation tasks (via translation, among others) to find out whether particular pragmatic functions, like supportive acts, are likely to take other linguistic forms or not. And, of course, the positive imperative and the hortative will be looked at in greater detail.

Notes

1 See, among others, De Rycker (1990: 44-90) and De Clerck (2006: 11-60) for more comprehensive accounts of the features and the subtypes of the English imperative.

2 This picture of the Dutch imperative will do here, but it is clearly not complete. In very formal contexts, for instance, imperatives such as (4a) and (4c) may take a *t*-ending. Another rare but interesting phenomenon is the existence of preterite and pluperfect imperatives in Dutch.

3 In his quantitative analysis, De Clerck (2006) makes a distinction between so-called major imperatives and minor imperatives like *listen*, *never mind*, *say* or *hang on* on the sole basis of their ICE-GB tagging as connectives, discourse markers or formulaic expressions. The usefulness of this dichotomy and in particular its operationalization are debatable. Figures 1 and 2 therefore present the frequencies of the major plus the minor imperatives, i.e. all formal imperatives in the corpus. Similarly, the numbers for Dutch include the differently parsed 'minor' imperatives in the CGN too, like *zeg maar* 'say', *kijk* 'look' or *kom* 'come on'.

4 The plays are: *Joking Apart* (Ayckbourn 1978, 1980), *Summer* (Bond 1981, 1983), *Other Places* (Pinter 1981, 1984), *The Real Thing* (Stoppard 1982, 1984) and *Pravda* (Brenton and Hare 1985, 1986), translated, in that order, by Hoeksema, Sternheim, Alphenaar, Kouwenaar and Nijmeijer.

5 The plays are: *Souvenirs* (Lemmens 1974, 1978), *You Are My Mother* (Admiraal 1984, 1995), *Scratch* (Herzberg 1989, 1995), *The Buddha of Ceylon* (De Boer 1991, 1997) and *Blowing* (van den Berg 2003, 2004) translated, in that order, by Wagenaar, Holland-Cunningham, Rudge, Couling and Vergano.

6 For the sake of completeness: De Clerck (2006) also mentions discourse-related and non-directive functions, but they are so marginal that they are simply classified as miscellaneous, together with the indeterminate cases.

7 Gerunds like *no talking!* seem to approximate the infinitives in Dutch. The distribution, the uses and the history of these gerunds and their comparison with the Dutch infinitives is an interesting area for further research.

Primary sources

Admiraal, J. (1982), *U bent mijn moeder*. Amsterdam: International Theatre Bookshop.

Admiraal, J. (1995), *You Are My Mother*. Translation C. Holland-Cunningham. Amsterdam: International Theatre and Film Books.

Ayckbourn, A. (1979), *Joking Apart and Other Plays*. London: Chatto and Windus.

Ayckbourn, A. (1980), *Zonder gekheid*. Translation P. Hoeksema. The Hague: Haagse Comedie.

Bond, E. (1982), *Summer and Fables. With Service, a Story*. London: Methuen.

Bond, E. (1983), *Zomer*. Translation J. Sternheim. Amsterdam: International Theatre Bookshop, Publiekstheater.

Brenton, H. and D. Hare (1985), *Pravda. A Fleet Street Comedy*. London: Methuen.

Brenton, H. and D. Hare (1986), *Pravda*. Translation P. Nijmeijer. Amsterdam: International Theatre Bookshop, RO Theater.

De Boer, L. (1991), *De Buddha van Ceylon.* The Hague: Toneelgroep De Appel.
De Boer, L. (1997), 'The Buddha of Ceylon'. Translation D. Couling, in: D. Couling (ed.) *Dutch and Flemish Plays*. London: Nick Hern. 3-61.
Herzberg, J. (1989), *Kras*. Amsterdam: International Theatre Bookshop, Toneelgroep Amsterdam.
Herzberg, J. (1995), *Scratch*. Translation J. Rudge. Amsterdam: International Theatre and Film Books.
Lemmens, G. (1974), *Souvenirs*. Amsterdam: Toneelgroep Centrum.
Lemmens, G. (1978), *Souvenirs*. Translation M. Wagenaar. Budapest: Centre Hongrois de l'I.T.I.
Pinter, H. (1982), *Other Places. Three Plays*. London: Methuen.
Pinter, H. (1984), *Vier eenakters*. Translation C. Alphenaar. Amsterdam: International Theatre Bookshop, Toneelgroep Centrum.
Stoppard, T. (1982), *The Real Thing.* London: Faber and Faber.
Stoppard, T. (1984), *Ware liefde.* Translation G. Kouwenaar. Amsterdam: International Theatre Bookshop.

References

Aijmer, K. and B. Altenberg (1996), 'Introduction', in: K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in Contrast.* Papers from a symposium on text-based cross-linguistic studies. Lund: Studentlitteratur. 11-16.
De Clerck, B. (2006), *The Imperative in English. A Corpus-based, Pragmatic Analysis*. Unpublished PhD thesis, Ghent University.
De Rycker, T. (1990), *Imperative Subtypes in Conversational British English. An Empirical Investigation*. Unpublished PhD thesis, University of Antwerp.
Dutch Language Union (2004), *Corpus Spoken Dutch*. Release 1.0. The Hague.
Geukens, S. (1986), *Sentence Type and Illocutionary Force. A Study of the Semantics of the Traditional Categories of Sentence Types*. Unpublished PhD thesis, University of Leuven.
Hopper, P.J. and E.C. Traugott (1993), *Grammaticalizaton*. Cambridge: Cambridge University Press.
Johansson, S. (1998), 'On the role of corpora in cross-linguistic research', in: S. Johansson and S. Oksefjell (eds.) *Corpora and Cross-Linguistic Research. Theory, Method and Case Studies*. Amsterdam: Rodopi. 3-24.
Mauranen, A. (2002), 'Will 'translationese' ruin a contrastive study?', *Languages in Contrast*, 2: 161-185.
Nelson, G., S. Wallis and B. Aarts (2002), *Exploring Natural Language. Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
Survey of English Usage (2006), *International Corpus of English. The British Component.* Release 2. London.
van der Auwera, J. (2006), 'Imperatives', in: K. Brown (ed.) *Encyclopedia of Language and Linguistics*, 2nd edition. Oxford: Elsevier. 565-567.

van der Auwera, J. and M. Taeymans (2004), '*Let's*, in English and in Dutch', in: G. Bergh (ed.) *An International Master of Syntax and Semantics*. Papers presented to Aimo Seppänen. Göteburg: Universitas Gothoburgensis. 239-247.

Van Olmen, D. (2008), 'Looking at *look* and *kijk*. A corpus-based, contrastive approach to the rise of discourse markers out of imperatives of perception in English and Dutch'. Paper presented at *New Reflections on Grammaticalization 4*, July 16-19, Leuven.

Part III

Corpus compilation, fieldwork and parsing

Caribbean ICE corpora: Some issues for fieldwork and analysis[1]

Dagmar Deuber

University of Freiburg

Abstract

In the Caribbean, English forms the upper segment of speech continua ranging from the Standard to the broadest Creole of each territory; social and stylistic factors correlate with the linguistic range. This paper explores the implications of this for the Caribbean components of the International Corpus of English *(ICE). The first issue addressed is how the most informal category of texts that field-workers are required to record for the corpus, conversations, can be made to fit into the segment of the continuum that can be described as English. It is shown that a compromise between the demands of recording 'English' and recording 'conversations' can be reached. The paper then goes on to discuss analytical approaches to grammatical variation in the Caribbean ICE corpora, demonstrating that the data can be fruitfully examined by a combination of quantitative and discourse analytic methods where corpus linguistics is closely integrated with sociolinguistics.*

1. Background and issues to be addressed

The *International Corpus of English* (ICE) lays down a common design for all individual components, which include corpora representing countries where English is a native language (ENL), such as Great Britain or New Zealand, as well as countries (or regions) where it is a second, official language (ESL), for example India or East Africa. Each component is to consist of 300 spoken as well as 200 written texts of approximately 2,000 words each, belonging to a variety of text categories. The composition of the spoken component, which this paper is exclusively concerned with, is shown in table 1.[2]

A problem in ESL contexts is that some of these categories of texts and in particular the private conversations which make up such a large part of the spoken component can be difficult to obtain. As Schmied has pointed out:

> In ESL communities English is normally used in domains related to the upper part of the formal spectrum. It may therefore be difficult to find texts for the spoken private categories, because other languages are preferred in conversations among family members and friends. […] The vast majority of the direct conversations in ICE-GB would

> simply not be conducted in English: all the family conversations (e.g. S1A-007) and mealtime conversations (e.g. S1A-056) in the British corpus would be too exceptional to be included in an African or Asian corpus of English. (Schmied 1996: 185-186)

Due to this difficulty, the East African component of ICE actually contains only 30 (instead of 90) private conversations and these required different strategies to record than in the ENL context (cf. also Holmes 1996: 169-170 on recording maximally 'natural' conversations for ICE New Zealand); they are discussions between students recorded in the classroom or conversations between Kenyans and foreign fieldworkers (cf. Hudson-Ettle and Schmied 1999).

Table 1: ICE spoken text categories (number of texts in brackets) and text codes

Dialogues (180)	Private (100)	Conversations (90)	S1A-001 to 090
		Phonecalls (10)	S1A-091 to 100
	Public (80)	Class Lessons (20)	S1B-001 to 020
		Broadcast Discussions (20)	S1B-021 to 040
		Broadcast Interviews (10)	S1B-041 to 050
		Parliamentary Debates (10)	S1B-051 to 060
		Cross-examinations (10)	S1B-061 to 070
		Business Transactions (10)	S1B-071 to 080
Monologues (120)	Unscripted (70)	Commentaries (20)	S2A-001 to 020
		Unscripted Speeches (30)	S2A-021 to 050
		Demonstrations (10)	S2A-051 to 060
		Legal Presentations (10)	S2A-061 to 070
	Scripted (50)	Broadcast News (20)	S2B-001 to 020
		Broadcast Talks (20)	S2B-021 to 040
		Non-broadcast Talks (10)	S2B-041 to 050

(adapted from www.ucl.ac.uk/english-usage/ice/design.htm)

In the anglophone Caribbean countries of Jamaica, and Trinidad and Tobago (T&T), for which ICE corpora are also being compiled,[3] English coexists with English-based Creoles and therefore has a status in between a native and a second language (cf. e.g. Mair 2007). Görlach (1991) has referred to this type of situation as "English as a second dialect" (ESD). Such situations present problems for ICE compilers which are, to some extent, similar to those faced by researchers working in ESL contexts. In the Caribbean, too, English is not generally the language of choice in conversations with family members and friends, and some of the conversations included in these two corpora were recorded using similar strategies as those employed in the East African context, i.e. setting up discussions or interviews in semi-formal contexts or having a foreign fieldworker present (cf. Deuber 2009b; Youssef and Deuber 2007). However, the situation in the anglophone Caribbean also poses additional challenges, both for fieldwork

and analysis, because the situation is not a dichotomous one, as in ESL contexts, but one where Standard English and the related Creole form opposite poles of a continuum of language use, which means that the Creole will have to be integrated into these corpora in a way that unrelated languages in the ESL context cannot. As a first illustration of this continuum of language use, consider the following example adapted from Allsopp (1996: s.v. creolized English):

(1) ai tould [h]im
ai told [h]im
a told im
a tol im
a tel im
mi tel im
mi tel am

The italicized versions represent what Allsopp describes as Creolized English. In contrast to the English versions above those in italics, which allow only certain Caribbean accent features, Creolized English may have Creole-influenced morphosyntactic features such as unmarking of past reference verbs (*tel*), while overt Creole features such as distinctive pronoun forms (*mi*, *am*) are confined to the lowest reaches of the continuum, according to this example.

When one considers whole texts rather than isolated sample sentences, the issue becomes even more complicated. In all but the most formal contexts there is likely to be some variation between different levels of language use. Mair (2002: 36) has described educated spoken English in Jamaica as a variety which "comprises an upper-mesolectal range on the continuum, and additionally allows for occasional forays into more basilectal territory".[4] This means that Creole or basilectal forms of the type illustrated in (1) above by the pronoun forms *mi* and *am* are not necessarily excluded from a text whose language would overall be classified as English or Creolized English, though they will not occur in large numbers. Moreover, text frequency is crucial even for features consisting in "morphological and syntactic reductions of English structure" (Allsopp 1996: lvi), e.g. unmarking of past reference verbs as illustrated in (1), which are considered characteristic of Creolized English or "informal" English usage (Allsopp 1996: lvi). In his detailed study of the intermediate range or mesolect in Jamaica in the framework of quantitative sociolinguistics, Patrick (1999: 201) divides his speakers into three groups, a 'High' one in whose speech only 29 percent of past reference verbs are unmarked, a 'Middle' one with a rate of unmarking of 64 percent and a 'Low' one with a rate of unmarking as high as 90 percent. This grouping can be correlated with the social status of the speakers (cf. Patrick 1999: 288). The policy of ICE is to include only 'educated' speakers (cf. Greenbaum 1996: 6) and one might expect that this would automatically restrict the degree to which Creole forms are used, but in addition to speakers' social characteristics, style is an important factor that has to be taken into account. This is illustrated in figure 1 by Winford's (1980) findings on past marking from his

quantitative sociolinguistic study of an urban community in Trinidad. His study included four social classes, I being the highest and IV the lowest. Sociolinguistic interviews were conducted for all social classes. Speakers of social classes III and IV were also recorded in informal peer-group interactions.

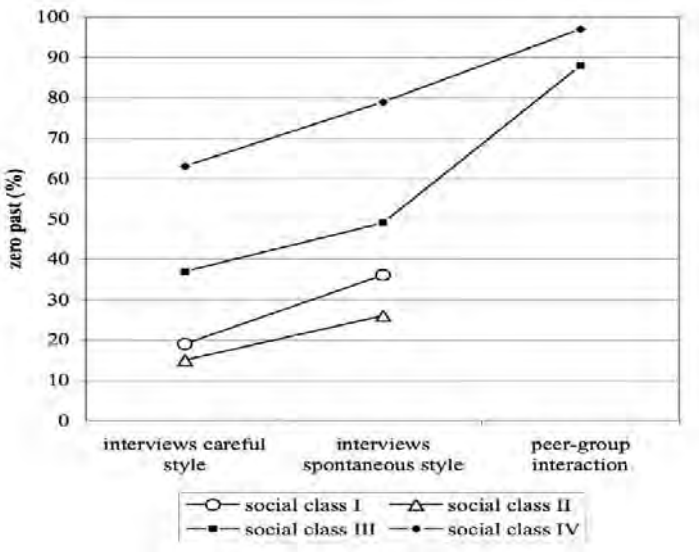

Figure 1: Past marking across the Trinidadian continuum[5] (based on Winford 1980: 57)

One can see in figure 1 that unmarking of past reference verbs increases from careful to spontaneous interview style for all social classes. What is also apparent is that the classic sociolinguistic interview can capture only a middle range of the continuum. This is as expected, since this situation, as Winford (1972: 161-162) remarked, "most likely does not encourage either the formal speech of, say, public address on the one hand, or the casual conversation common among friends and family members on the other".[6] The difference between even the spontaneous interview style and peer-group interaction is quite pronounced for this variable, especially for social class III. This social class almost converges with class IV on a very low level of past marking in the peer-group context. In contrast, English varieties with a very high level of past marking are conspicuously absent, as Winford did not extend the range of the sociolinguistic interview in that direction. This is, of course, precisely where ICE comes in: given the preponderance of public/formal contexts in the design for the spoken ICE component, the Caribbean corpora can certainly be expected to cover the highest portion of the continuum. The special challenge for fieldwork for the 'private' interactions is to find an adequate compromise between the somewhat conflicting

demands of recording 'English' and recording 'conversations', targeting an appropriate level of language use in a situation where the boundaries between English and Creole are fluid. Also, if English, as represented in the Caribbean corpora, comprises the whole upper portion of the continuum and may incorporate shifts into its lower reaches, analysts have to take account of this variability, which raises further issues. It has already become clear in the discussion so far that frequencies of Creole features play an important role in defining varieties along the continuum, but how far can these take the analyst in accounting for the variation observed? Quantitative studies of Creole continua are certainly open to the same kind of criticism as quantitative sociolinguistic studies generally:

> It is entirely reasonable to assume that there will be some tendency for higher frequencies of use of a sociolinguistic variant to be more robust carriers of that feature's indexical meaning than lower frequencies. But to map a linear dimension of "social meaning intensity" precisely onto numerical arrays is extremely counter-intuitive. This procedure side-steps a theory of sociolinguistic indexicality which […] gives priority to the local contextualisation of single variants in discourse. (Coupland 2007: 76)

The present paper is centrally concerned with these analytic issues. In section 3, illustrative findings from ICE-Jamaica and ICE-T&T for morphological and syntactic Creolisms will be presented and discussed from this perspective. First, however, I will elaborate on the issue of fieldwork (section 2) as the basis for the kind of data to be analysed. The final section (4) will present some brief conclusions on the implications of Caribbean Creole continua for the compilation and analysis of ICE corpora.

2. Fieldwork: recording conversations

The fluid boundaries between English and Creole are, of course, an issue not only in the case of conversations. Mair (1992) comments as follows on data from a Jamaican phone-in programme which he analysed as part of his "pilot study" of "problems in the compilation of a corpus of standard Caribbean English":

> This is definitely not international standard English. Nor is it likely to be accepted as Jamaican standard by local speakers. But the text was produced by an educated speaker in the relatively public and formal context of a radio programme, and one may well ask whether the concept of the educated standard the compilers of ICE seem to have in mind is adequate to a description of spoken data from the Anglophone Caribbean. (Mair 1992: 93)

However, in the case of radio programmes and other public text types, fieldworkers are limited to making an appropriate selection from a given range of texts, whereas in the case of conversations they can influence the recording situations. The following remarks on fieldwork are therefore concerned specifically with the recording of conversations.

In the case of ICE-Jamaica, the majority of recordings in the category 'conversations' were made by graduate students at the University of the West Indies, Mona. For the most part, they tried to engage other members of the university in a conversation in a variety considered as English, though they were not supposed to intervene if the language changed in the course of the conversation, nor were recordings excluded in the process of corpus compilation on the grounds that Creole was used. This approach met with the expected difficulties, as one of the fieldworkers relates in one of the texts; the relevant extract is printed below as (2). In this example and subsequent ones from the corpora, the text has been simplified by leaving out some of the backchannels and hesitations in order to increase readability. In order to follow the examples it must be noted, however, that in the case of overlaps, marked by <[> … </[> for each string involved in the overlap and <{> … </{> for the whole set of overlapping strings, other speakers' overlapping strings – numbered if there are several – are always given at the end of a complete speaker turn (see e.g. A's last turn in (2)). Other markup symbols that appear in the examples include: <$>, the speaker identification symbol; <#>, which indicates the beginning of a text unit roughly corresponding to a sentence in written English; and <,>, which marks a short pause (for further details of the markup, see Nelson 1996, 2002).

(2) <$A><#>And I mean trust me it has not been easy trying to get the data like you go to persons and you're like <#>Alright would you like <unclear>word</unclear> to have a conversational interview for like half an hour <#>So long <#><{><[>That's the <?>different</?> thing</[>
<$B><#><[>Ah yeah <#>Uh well the thing is</[></{> half an hour isn't actually long but when you sit down on a table and it's like timing yourself and talking into a little recording device then it becomes long and it just <#>It feels tedious
<$A><#>Right and like when they hear that it will be recorded<,> worse in <{><[>English cos</[>
<$B><#><[>Yeah<,> and</[></{> oh yeah for true cos yeah most people want to break down after a while <#>I mean it is hard to have a conversation with another Jamaican under non-academic <{><[>situation<,></[> and speaking Standard English that's really hard
<$A><#><[>That's right<O>laughs</O></[></{>
<$A><#>Yeah <#>And that kind of slows down getting the data<,> because <{1><[1>you know</[1> it should be informal and then you know after that <?>you have to tell you that</?> informal and then just <{2><[2>chatting away<,></[2> <#>Mhm
<$B><#><[1>I can imagine that it's</[1></{1>

> <$B><#><[2>That's a problem that's a problem</[2></{2>
> <$B><#>Especially if you never grew up in an environment where you were exposed to Standard English every day <#>Standard English became something that you had to do only for academic purposes (ICE-Jamaica S1A-030)

In spite of these difficulties, the fieldworkers were able to obtain a large number of recordings. As the detailed analysis in Deuber (2009b) shows, these provide some interesting examples of code-switching into Creole but are overall best described as representing 'informal Jamaican English', and are thus entirely appropriate for the corpus and the text category. Of course, the texts are not uniform. The analysis in Deuber (2009b) includes forty texts (S1A-001 to 040) and these were found to range from semi-informal, interview-like ones to truly informal, conversational ones, which tended to show more Creole features. The following factors were identified as relevant for variation between the texts:

- relation between participants (friends, acquaintances, strangers);
- setting (e.g. office, library etc. versus outside, in a hall of residence etc.);
- topic (e.g. studies/work versus more personal topics);
- speakers' individual backgrounds and preferences (there is a minority of Jamaicans for whom English is a home language, while for the majority it is not, as also noted by speaker B at the end of (2) above);
- fieldworker (although the majority of the texts were recorded by Jamaican fieldworkers, a few were recorded by non-Jamaicans).

In the case of ICE-T&T, 17 of the 20 conversations included in the corpus so far (S1A-001 to 017) were recorded by students in the final-year linguistics research class at the University of the West Indies, St. Augustine, among teachers in secondary schools in Trinidad (for details of this project, see Youssef and Deuber 2007). The remaining three (S1A-018 to 020) are actually conversations between a fieldworker and friends of his, with a mixture, however, of general or current affairs topics and a few more personal topics. The school recordings represent a range similar to the one found by Deuber (2009b) in the ICE-Jamaica conversations, in this case from discussion-type texts, as there are always at least three participants, to more conversational ones. Two important factors for variation in this sample seem to be the topic and the relationship between the fieldworker and the other participants, i.e. whether the fieldworker was an outsider or an insider, as happened in some cases where a student also worked as a teacher in a school, or a teacher took on the task of recording his or her colleagues for a student. Of course, these two factors are not unrelated, as the relationship between the speakers and the atmosphere that is established in their interaction is likely to influence the course of a conversation, and, even if the topic is set, it can be treated in quite different ways, i.e. in a more objective or more personal manner. In addition to these factors, individual backgrounds and preferences are as relevant in T&T as in Jamaica. As Youssef put it:

> For the majority […] the Creole is simply their more comfortable variety, and they assign value to Standard English as the code of education and informed discussion, of public as distinct from private address. Again, for those for whom Standard English is their first and most comfortable variety, these social functions will be less determinant of their use of it. In practice, there are a myriad of social factors to which an individual is attuned in making his/her language choices. (Youssef 2002: 19)

In sum, although recording appropriate private interactions for Caribbean ICE corpora is certainly a challenge, fieldworkers have so far been able to strike a good balance between the demands of recording 'English' and recording 'conversations'. The recordings represent a range of language use, which is determined by a complex interplay of situational and social factors.

3. Analysis: discussion of selected findings for morphology and syntax

This section will consider by which methodology the Creole continuum, as reflected in Caribbean ICE corpora, is best approached. Specifically, it will address the issue of what a quantitative correlational approach, as illustrated by the data from Winford (1980) in figure 1, can show and what could be additionally gained by a more interpretive approach of the type that has recently become prominent in the sociolinguistics of style.

As Deuber (2009b) offers comprehensive analyses of selected ICE-Jamaica data from both quantitative and qualitative perspectives whereas in the case of ICE-T&T only findings for a few selected features and texts have been previously presented (Deuber and Youssef 2007; Deuber 2009a), this section will focus on data from ICE-T&T, though some of the Jamaican data will be revisited as well.

At the time of writing, the following texts in ICE-T&T have been completed, while many others are at various stages of transcription and correction: 20 conversations, as already mentioned in section 2 (S1A-001 to 020), the complete category 'class lessons' (S1B-001 to 020), 15 unscripted speeches (S2B-021 to 035) and the complete category 'broadcast news' (S2B-001 to 020). These are the texts used for the analyses that will be discussed here.[7]

3.1 Variation across text categories

Figure 2 combines Winford's (1980) findings for past marking with the new data from ICE-T&T.[8] Figure 3 shows the same full range of variation for another variable, use versus non-use of the auxiliary in present-tense progressive forms.[9]

The results for both variables are very similar in several respects. The frequencies of the Creole variants are in the range of zero to five per cent in the two monologic text categories from the ICE corpus, with only a minor increase from the scripted broadcast news to the unscripted speeches. In the two dialogic text

categories from the ICE corpus the frequencies of the Creole variants are considerably higher than in the monologic ones. The difference between the class lessons and conversations is in both cases only very minor. One would expect, however, that other public dialogues will occupy the range in between the monologues and dialogues included so far, as these are broadcast or involve more formal settings, and, in contrast to class lessons, normally have only adult participants.[10] A further similarity between the results for the two variables is that the aggregate figures for ICE conversations are roughly similar to those for Winford's two higher classes in the careful style in the sociolinguistic interview.

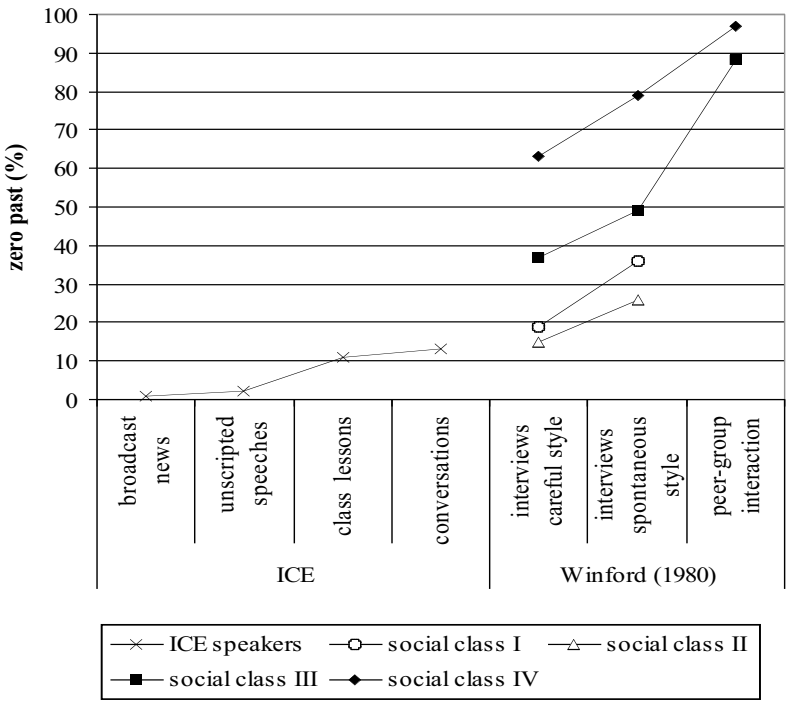

Figure 2: Past marking across the Trinidadian continuum II

As was already noted in the discussion of figure 1, unmarking of past reference verbs is considerably more frequent in spontaneous than in careful interview style. Figure 3 shows that this applies to non-use of the auxiliary in present-tense progressive forms as well, and, again, the highest frequencies of Creole features are reached only in the peer-group interactions recorded for the two lower social classes. A difference between the results for the two variables is that zero copula in present-tense progressive forms is considerably more frequent in the ICE dialogues as well as Winford's interview data than unmarking of past reference verbs.

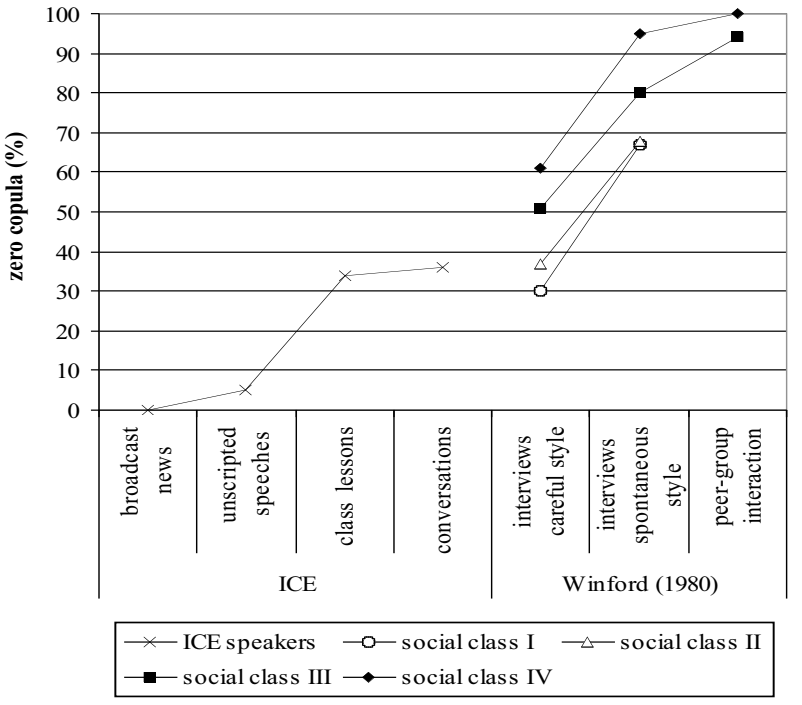

Figure 3: Present-tense progressive forms across the Trinidadian continuum (partly based on Winford 1980: 57)

Overall, one can say that the Caribbean ICE corpus considered here represents the upper portion of the Creole continuum very well. This can be seen clearly in the way each of the Creole features considered in this section is distributed across the different text categories. In addition, the results point to differences between features, as expected in a continuum situation where features are not simply 'Creole' but 'more or less Creole'. Such differences will be explored further in the next section.

3.2 Comparing different Creole features within a text category

Figure 4 displays the frequencies of selected Creole features in conversations in both of the Caribbean ICE corpora. The figure includes four features involving non-use of inflections or function words required in Standard English and four overt Creole forms. Unmarking of verbs is shown for past tense as well as third-person singular present-tense forms, and non-use of the copula for progressives and predicative adjectives. The four overt Creole features are *me* as a subject pronoun (unstressed), as illustrated by example (1) in section 1, the same form as

a possessive pronoun, a further Creole pronoun form, *them* as third-person plural subject pronoun, and *ain't* and *no* as markers of verbal negation in T&T and Jamaica, respectively. It must be noted, however, that two of the pronoun forms, namely *me* and *them* as subject pronouns, are extremely marginal in Trinidad even in the Creole,[11] which is not the case in Jamaica.

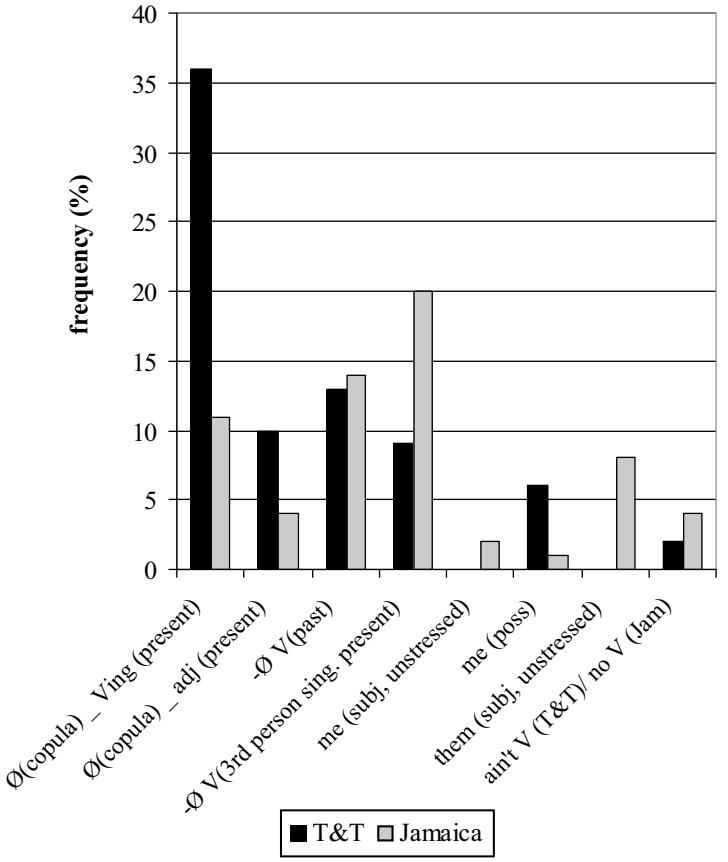

Figure 4: Frequency of selected Creole features in conversations in Caribbean ICE corpora (ICE-T&T S1A-001 to 020 and ICE-Jamaica S1A-001 to 040)

Figure 4 shows that, in both corpora, there are important differences between the frequencies of different Creole features, which can be related to different degrees of 'Creoleness', since the more strongly a feature is associated with the Creole, the less likely it is to be used in the kind of data analysed here. The tendency is

for zero forms to be more frequent than overt forms but there seems to be a hierarchy among both (with differences between the T&T and the Jamaican data which are beyond the scope of the present paper to explore).[12]

3.3 Intertextual variation within a text category

In the analysis of the first 15 texts in the category 'class lessons' in the ICE-T&T corpus presented in Deuber (2009a), the present-tense progressive variable was not only analysed for the whole sample, as in section 3.1 above, but also by individual texts. Though the different social and situational factors that might influence language use in each particular text cannot really be disentangled in such a small sample, it did appear that the Creole variant without the copula tended to be favoured more in lessons in science, technical and vocational subjects as opposed to humanities subjects. Figure 5 shows the same type of analysis for the seventeen conversations in ICE-T&T recorded in the school setting (cf. section 2). Absolute figures are given here, as percentages have to be taken with caution where a text has a rather low overall frequency of the variable. The topic as one factor likely to influence language use has been included for each text. The data do not allow simple correlations between topics and the frequency of the Creole variant, of course (cf. e.g. the different results for the many texts with the topic 'language use in the schools'), and other factors must be considered as well. When additionally taking into account the factor of fieldworker status (which is not known exactly for all texts, however), what one can say is that it is often in texts in which personal matters are discussed and/or the fieldworker is an insider that the frequency of the Creole variant is high. For example, in texts S1A-008 and S1A-011 both applies, i.e. personal matters are among the topics discussed and the fieldworker is an insider. In text S1A-007, the same set topic as in several other texts, 'language use in the schools', is discussed, but the recording was made by a teacher who seems to know quite well the other two teachers who participate, which is not the case in some of the other texts with this topic.

Of course, in order to better assess the degree of Creole use in individual texts, one would have to look at more than one variable, but this is hampered by the fact that many grammatical variables are too infrequent to be meaningfully analysed in a 2,000 word text. Furthermore, the occurrence especially of Creole features which are rather infrequent overall may be due to style-shifts within a text, which need to be looked at in context rather than by comparing frequencies across texts. A case in point is *does* as a marker of present habitual aspect in Trinidad. This is not only very rare in the ICE data, with a total of only 18 occurrences in the 20 conversations included so far and as few as eight in the 20 class lessons, but if it occurs it is mainly in special contexts, for example in the conversations when speakers emphasize a point in connection with the identities being projected in discourse (cf. example (8) in section 3.4 below; also example (9)) or in the class lessons in personal remarks to the students (Deuber 2009a). The present-tense progressive form in the ICE-T&T conversations is probably the variable best suited for the kind of analysis presented in this section, as the form

is relatively frequent overall and the frequency of the Creole variant is high as well. However, even for such a form the lower limit for a quantitative variable analysis is normally a whole text. An analysis of intratextual variation will have to consider how Creole features are distributed in a text, without, however, analysing each of them quantitatively, and especially what their function and meaning is in context. This will be illustrated in the following section.

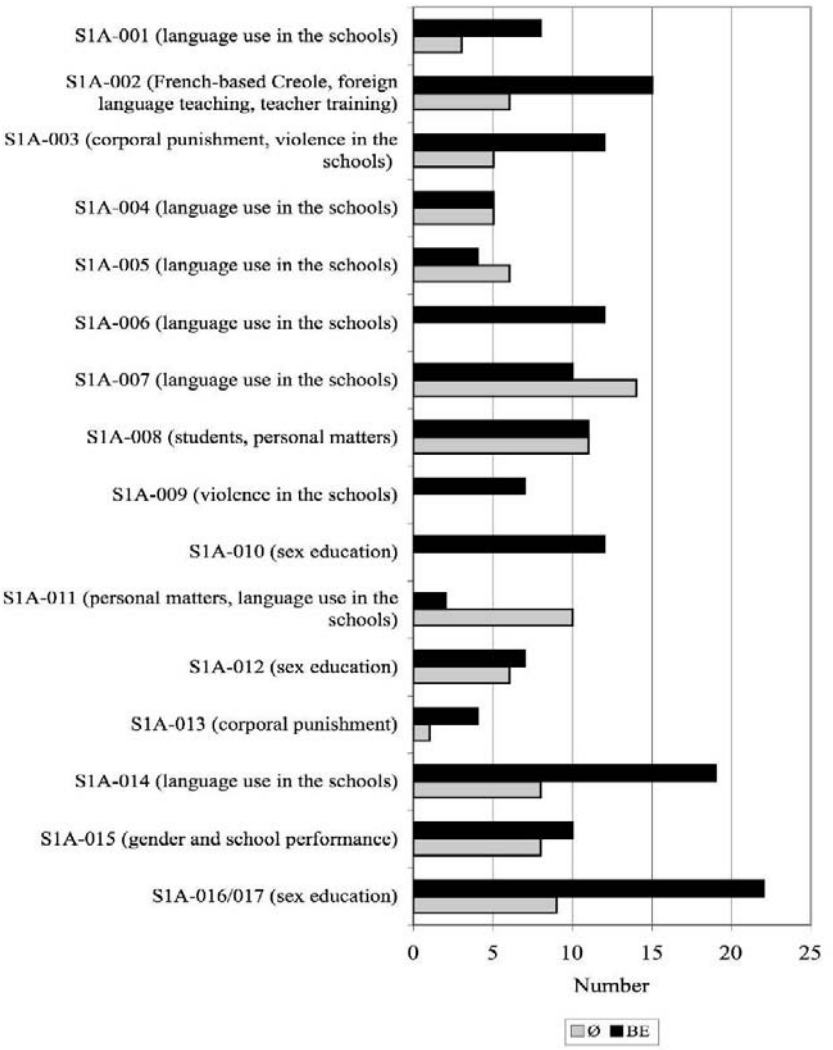

Figure 5: Present progressive forms with BE versus Ø in conversations in ICE-T&T (S1A-001 to 017) by individual texts (topics of texts in brackets; texts listed together are from the same recording)

3.4 Intratextual variation

As a first illustration, consider the three examples below. Creole forms are highlighted in italics, as will be the case in all examples in this section (an overview of the relevant morphological and syntactic features of Jamaican and Trinidadian Creole is provided in the appendix).

(3) They still have to start from scratch with that because what you have happening with the language arts is the nineteen eighty-six <#>Tell me about *me* time eh <#>Nineteen eighty-six I'll tell you about change (ICE-T&T S2A-029)

(4) And so she proposes to build six hundred or find spaces in six hundred pre-schools across the country by two thousand and ten<O>member-of-audience-sucks-teeth</O> <#>That has <.>start</.><O>audience-laughter</O> <#>That has started<O>audience-laughter</O> <#>But that came with some concern as we just heard<O>audience-laughter</O> <#><&>low-voice</&>I find *all you dissing* me <un-clear>words</unclear><&>end-of-low-voice</&><O>audience-laughter</O> <#>But that came with some concern in that if you are going to be uhm responsible for the foundation years the formative years of these children five O levels and a certificate course just really wouldn't cut it (ICE-T&T S2A-030)

(5) Cos I mean a thing like the Constitution <,> <#>Even *me no* understand someone who knows English <#>Cos I mean there's a problem in terms of rights <#>A Creole speaker cannot understand so what about him (ICE-Jamaica S1A-040)

Examples (3) and (4) are from texts in the category 'unscripted speeches' in the ICE-T&T corpus. The texts are from two different guest lectures in an education class at the University of the West Indies. In both examples, a Creole pronoun form is used, possessive *me* in (3) and the second-person plural pronoun *all you* in (4), which is a rare and marked form in the ICE-T&T data (cf. Deuber 2009a); (4) also contains one of the few present-tense progressive forms without auxiliary BE in this category of texts. What is noteworthy here is that the text units containing these forms are clearly separate from the lecture itself. In (3) the speaker briefly interrupts the lecture to ask the organizer to remind her of the time available, while in (4) the speaker is briefly addressing the audience to complain that something she has said is not being taken seriously.

Example (5) is from a conversation, this time in the ICE-Jamaica corpus. The text is one of the interview-type ones and does not have many Creole features overall. The example shows two overt Creole features which are generally rare in the data (cf. figure 4 in section 3.2) being used in what may be described as a symbolic act of identity with Creole speakers not proficient in English. What is important in comparison to examples (3) and (4) is that they are used while the

speaker is making a point within the ongoing discourse, rather than stepping outside of it.

It is generally in the conversations that Creole forms are most integrated into the texts, and it is here that their functions are most varied. While the analyses in section 3.1 above have indicated that the class lessons in ICE-T&T are similar to the conversations in terms of the overall frequency of Creole forms, the use of many Creole features, except the most frequent ones like Ø *-ing*, seems to be mainly related to the opposition between personal remarks to students and regular classroom discourse, as Deuber (2009a) has shown. In the conversations, the situation is somewhat similar as more personal topics or viewpoints certainly favour the use of Creole, but such a generalization cannot adequately capture the many shades of contextual meaning that Creole forms may have in conversations. This will be illustrated below by examples from three conversations, one from ICE-Jamaica and two from ICE-T&T.

The conversation from ICE-Jamaica that will be considered here stretches over three texts in the corpus, S1A-018 to S1A-020, which are all based on the same recording. Examples (6) and (7) are from the first and last text in this group, respectively. The recording is very much in the interview format. The interviewer (identified as A in (6) and as B in (7)) asks a question and then limits herself to backchannels (largely omitted here) as the interviewee responds.[13] Example (6) illustrates this pattern very well. The example shows the beginning of the text, where the interviewer asks the first question. There is only a very short interruption in the interviewee's extended answer, where he asks the interviewer if she knew that Tallahassee was the capital of Florida. The example is also a typical one for the recording in terms of the topic – the recording is mainly about the interviewee's work and career – as well as in terms of language: in much of the group of texts there are only very few, isolated Creole features.

(6) <$A><#>We're speaking for twenty minutes <#>I'm just going to ask you general questions <#>It really does not matter what we speak about <#>But since you are at Carimac[14]<O>speaker-B-starts-laughing</O> I'll ask you about job-related <{><[>things that</[>
<$B><#><[>Oh <?>you</?> really</[></{> can talk to me about me too <#>I'm trying to put my phone on silence here <#>Go on <{><[><,>so ask your questions then at all</[>
<$A><#><[>So uhm just tell me</[></{> how long you've been working here what your job is about what are the most interesting things about your job
<$B><#>Wow that's a lot <#>Let's go one question at a time<O>laughs</O> <#>I'm putting the phone on silent here so it's not ringing while you are<,> recording <#>So it's on silent thank God <#>So twenty minutes starting now <#>I'm too strict on time you know <#>I'm learning to manage my time that is the thing I'm doing right now <{><[>but <?>good</?></[> <#>It is a good thing <#>But let's go through your questions now <#>So the first thing is what is it I do here at

Carimac <#>I'm a lecturer now <#>I joined the staff here in July two thousand and one after coming back from the United States <#>I was away studying on a Fulbright scholarship <#>I did my Master's in Communication Research and Theory for two years at Florida State University in Tallahassee <#>You know that Tallahassee is the capital of Florida right
<$A><#><[>But that's a good thing</[></{>
<$A><#>I did not know that
<$B><#>I didn't know that either <{1><[1><,><#>When they</[1> told me that I'm going to the capital of Florida <?>I</?> *say* yeah Miami<{2><[2><,></[2> <#>And then they *say* no Tallahassee <#>But you <?>want</?> to think Miami is the capital of Florida <#>Anyway so I went to Tallahassee for two years starting in <}><-><.>Sep</.></-><+>September</+></}> in August nineteen ninety-nine came back in May came back exactly on May thirty-one two thousand and one and I was in touch with Professor Brown because there was an opening here at uhm Carimac a senior lecturer position in social marketing <#>So I took it upon myself to apply while I was still abroad studying you know to be <?>correct you know not</?> wait until I come back to the country to actually apply for a job while I'm back here <#>So I applied for it like January which was the January of the last semester I had at Florida State University <#>So that was January two thousand one I applied <#>May two thousand one I was in touch <#>I got back here on the thirty-first at five o'clock in the afternoon and Friday <#>And that was a Thursday <#>And Friday morning I was attending the department's retreat <#>So it was good it all worked out and so on <#>And the contract was sorted out in June May June July <#>And July one I started working <#>I had my contract and everything was good <#>So I've been here since <#>I started off as an assistant lecturer cos I had a Master's degree but based on my performance being a Fulbright scholar having published<,> having presented papers at conferences all of that stuff helped me to get a three-year contract because the way this university is organized you know they have different types of contracts <#>A lot of people get temporary appointments which is like a one-year contract <#>But when it's like a three-year contract it means that you're kind of good
<$A><#><[1>I thought it was Miami</[1></{1>
<$A><#><[2>Miami <#>Yes</[2></{2> (ICE-Jamaica S1A-018)

Example (7), by contrast, shows an exceptional part of the recording. Here the topic changes to a personal one, the interviewee's hobby of swimming, and there is a cluster of Creole or Creole-influenced forms and constructions in this segment.

(7) <$A> […] <#>In terms of what I do for non-academic work I try to swim every Sunday <#>Now I'm learning to swim<{1><[1><,> which is

good</[1> <#>I'm learning to swim <#>I didn't know to swim all these years <#>I had a <{2><[2>fear of water</[2> <#>Yes <#>*What that has to do* with it though<O>laughs</O> coming from <{3><[3>Kingston<O>laughs</O></[3>
<$B><#><[1>Oh <#>Yeah</[1></{1>
<$B><#><[2>You're from Kingston</[2></{2>
<$B><#><[3>Persons from Kingston</[3></{3> <?>that</?> can't swim
<$A><#>Really you think so no man people can swim <#>It's just that I had a fear <#>When my father *did* try to teach me people were laughing at me on the beach and this was at under five <#>And it was a traumatic experience<{1><[1><,> learning to swim</[1> on a crowded beach <#>Don't take your child to teach them to swim on a crowded beach <#>Anyway so I *get* over the fear <#>In fact I was going to swimming classes while I was at Florida State University and the response from everybody <#><&>imitation</&>You're from Jamaica<O>speaker-B-laughs</O> and don't know how to swim<&>end-of-imitation</&> <#>And so *why you're making* the assumption that everybody who *live* on an island *know* how to swim <#>Anyway <#>I *come* back here and I've been going to the UWI swimming pool every Sunday since December twenty-seven last year <#>After Boxing Day <{2><[2><,>I *get* up and went <#>No I've been</[2> doing the self-taught thing you see and *I taking* no swimming lessons <#>But I've been eavesdropping on people's swimming lessons <#>For the swimming instructor would be there teaching other people and I'm standing within earshot <#>Okay okay so you must do the <?>Valentine arch</?> to get the breast stroke<,> and I'm doing all of that to get breast stroke and I don't know how to do the breast stroke but for the life of me I haven't gone into the deep end of the pool as yet <#>*I still a little bit afraid* <#>The breathing is a challenge for me and I remember while learning to swim in Florida State University they taught me to breathe through my nose and somebody told me that that's for competitive swimming <#>But now I'm learning that you must breathe <{3><[3>through your mouth</[3> <#>So I'm listening to all the tips and you know everybody *come* in the water <#>You say ha I don't know how to swim and everybody *give* you a little tip <#><&>imitation</&>Oh swim <?>mind</?> this and <unclear>word</unclear> <#>Practice this practice this<&>end-of-imitation</&> <#>So I'm getting there <#>I was out for a month or two because I had other things to do on Sunday <#>I was off the island <#>But I'm back now <#>I started back last week Sunday and it's wonderful <#>I know to do the breast stroke <#>The next stroke to do now is the backs no <{4><[4>the freestyle<,></[4> with the flutter kick and the <unclear>word(s)</unclear> you know the jargon <?>and</?> the talking about this <#>And then the backstroke <{5><[5>and then the</[5>
<$B><#><[1>Yes <#>I can imagine</[1></{1>
<$B><#><[2>So you can be <?>in the swimming lessons</?></[2></{2>

<$B><#><[3>Through your mouth</[3></{3>
<$B><#><[4>Back <#>Okay</[4></{4>
<$B><#><[5>The backstroke</[5></{5> is the best one for somebody who's just learning to swim (ICE-Jamaica S1A-020)

This example has also been presented in Deuber (2009b) and was interpreted there as an instance of topic-related style-shift. However, the example could also be interpreted further, using Coupland's (2007: chapter 5) framework for the analysis of identity contextualisation processes. One could say that in (7) the genre frame shifts slightly, from 'interview with someone in their official capacity' somewhat in the direction of 'conversation between equals about everyday topics'. Although the part of the recording shown in (7) is also largely in the interview format, it is slightly more interactive than much of the rest of the three texts, with the interviewer interrupting the interviewee with a personal question – whether he is from Kingston – and making a few comments. Her comment at the end of the extract suggests that she is probably more competent in this field than he is, whereas much of the rest of the interview is about his field of competence. Actually competence as an aspect of the interviewee's projected identity seems to be important to the interpretation of the style-shift. While the format of the interview with him in his official capacity constrains him to show himself as competent in his professional domain, the shift away from this genre frame allows him to cast himself in a somewhat different light – ambitious, as in his professional domain, but kept back by his fear of water.

While the association between highly standard language and professional competence seems a natural one in the sociolinguistic situation of the Caribbean, matters are not always that simple and competence can be interpreted in different ways, as example (8), from one of the teacher conversations in ICE-T&T, will show.

(8) <$C> […] <#>You can't tell if the child will improve in algebra if you test trigonometry next day<,> <#>So I'm asking <quote>so did you all ask the examiners that</quote> <#>She *say* <quote>yes<,> but you know they and all know that they are dealing with outdated material but that is what they have to work</quote> <#>I *say* <quote>well that is stupidness</quote><,> <#>They are teaching us to be teachers and they can't teach properly <#>You don't say<,> you tell the teachers <quote>oh this is what we have to work with</quote><,> <#>You are getting a logical question answer logically <#><quote>Oh that is what it is so we have to work with that</quote> that is nonsense <#>That is why I wanted to go and do my Dip Ed to see if they *go* tell me the same thing<{><[><,></[> <#>No I *get* up there and *rough* up some of them<,> <#>You think I *fraid* them<,> making joke or what
<$Z><X><#><[>Yeah<O>laughter</O></[></{></X>
<$Z><X><#>But the people in Dip Ed haven't been in school<,> for the last at least ten years they haven't been in class</X>

<$C><#>No the only reason<,> the only I *ain't* go through with that Dip Ed is the air-condition you know<,>
<$Z><X><#>Oh you *get* into Dip Ed</X>
<$C><#>No<,> I *sign* up I got through for it right<,> but I went to make a arrangement to see how I could get past the air-condition if I could come to class like let's say for a little while and do the thing externally <#>*Them* lock up *themself* in some office and you have to talk through intercom and all this kind of crap
<$Z><X><#>Yeah<,> that's how they have you</X>
<$C><#>All kind of crap <#>I *start* to cuss the man and I *walk* out<,> <#>He *ain't* know *who cussing* him<O>laughter</O> <#>He *ain't* know *who cussing* him
<$Z><X><#>But the people in Dip Ed<,> have a set of theories and thing about education</X>
<$C><#>That's right
<$Z><X><#>And last time they taught was probably fifteen years ago</X>
<$B><#>That's the problem<,> there's the problem <{><[><unclear>words</unclear></[>
<$C><#><[>No well that's what she *tell* me</[></{> <#>*They aware* that their techniques are outdated
<$Z><X><#>Yeah</X>
<$B><#><unclear>words</unclear> I now *finish* Dip Ed you understand<,> <#>I now *finish* Dip Ed so I now now *gone* through the whole set of crap that they *does* do <#>Cause they *does* preach one thing<,> do something totally different <#>Because they were trying to encourage the visual and performing arts group<,> this whole concept of uhm integrating the curriculum<,><#> They don't integrate the curriculum <#>Now the problem is we always used to complain about that<,> is a bunch of old<,> wrinkled<,> gray<,> people
<$B><#>Strike that out the <unclear>word</unclear><O>laugher</O>
<$C><#>They won't know your voice <#>They won't know your voice
<$B><#>A bunch of old wrinkled gray people who<,> really and truly<,> only *concern* about <#>You know is university lecturers <unclear>word</unclear> really <?>only</?> *concern* about your pocket you want to have a easy way out<,> you understand<,> <#>You have not teached like you said<,> right<,> you have not teached for a while so you really *ain't* know what it is like out there<,> <#>*You comfortable* here coming and lecturing to me<,> day in day out<,> right<,> and *you comfortable* with saying one set of stuff but you *done trap* in your ways already<,> <#>Now we always used to complain about that (ICE-T&T S1A-002)[15]

There is a high density of Creole features in this extract and these include not only such forms as zero copula or unmarked past reference verbs but also some of

the more marked overt Creole features, for example the negator *ain't* (cf. section 3.2 above), *does* as a marker of present habitual aspect (cf. the discussion in section 3.3 above) and the only instance of *them* as a subject pronoun in ICE-T&T S1A-001 to 020.[16] The use of these forms is certainly partly connected with the fact that speaker C talks about his personal experience and how he opposed the system, but it is also important that the teachers consider themselves more competent in the practical aspects of teaching than university lecturers in education with their theories, so one relevant aspect here in terms of the identities being projected is the contrast between a practical versus a theoretical orientation.

As a last example of the importance of context for the analysis of Creole features in the texts, (9) shows the beginning of conversation S1A-007 in ICE-T&T. In the analysis of the progressive variable in section 3.3, S1A-007 was noted to be a text with a high frequency of the Creole variant of this variable, and the status of the fieldworker as an insider was mentioned as a factor probably influencing the kind of language being used.

(9) <$A><#>What kind of language you feel *you talking* <#>You feel *you talking* standard English or what
<$B><#>Boy me *I not talking* any standard English you know <#>I talk standard English when I have to talk standard English<,> when the situation *warrant* it<,> but me I *does* talk *me* Creole <#>I <?>is</?> Trinidadian *I always talking me* Creole<,> you understand
<$A><#>And when *you in front* of the children *and them* what do you do
<$B><#>It depends on the situation again<,> <#>It depends on the situation <#>Sometimes when need be I talk *me* standard English <#>If at times *we talking*<,> for them to understand and communicate once I could communicate with them I use the Creole <#>So it is back and forth
<$C><#><}><->I I I<,> I agree with that</-> <=>I agree with that</=></}> because when *we liming* look *we liming* here<,> it's broken English Creole *we talking* <#>But I agree when the situation and the need arises we would speak the proper English
<$A><#>You feel *teachers talking* generally in Trinidad talking standard English with the children *and them*
<$C><#>Sometimes<,> <#>It depends again on the situation<#> If is a informal<,> setting not the classroom lesson<,> then I *meself*<,> talk broken English with the children because that is how they understand<#> But if is in a lesson I speak the Standard English the proper English the textbook English (ICE-T&T S1A-007)

Example (9) confirms that the participants interpret the situation as an informal one, as one of the speakers even describes it as "liming", a Trinidadian expression meaning 'to relax and chat'. However, it also becomes clear that this is not the only relevant factor, as Creole is also used to emphasize a Trinidadian identity. One can see clearly here that the broad correlations between language use and situational factors which can be established in the quantitative framework are not

sufficient to explain the use of Creole forms, especially in the conversations, where they are used most freely and have the widest range of functions.

4. Conclusion

English in the Caribbean is centrally defined by the fact that it represents the upper portion of a continuum ranging from the most standard variety of English to the broadest Creole. Social factors influence language use along the continuum but situational factors are very important as well. The ICE corpora are restricted to the language use of 'educated' persons but the spoken components, which this paper has been concerned with, comprise a broad range of situations from very formal, such as news broadcasting, to very informal, namely the private conversations which make up a considerable part of the spoken component. In the context of Caribbean Creole continua such as those in Jamaica and T&T, following this design in the way it has been done in Great Britain or New Zealand, for example, would lead to the inclusion of an inappropriately broad range of language use for a corpus of English. The solution that has been adopted by fieldworkers is to record interactions in semi-formal contexts for the conversation category. This means that some of the texts are more like interviews or discussions but, due to the multiplicity of factors that may influence language use in such broadly defined situations, there is inevitably a range of interactions from more formal to less formal ones.

Quantitative analysis shows that the texts that have been recorded in Jamaica and T&T for the category 'conversations' can be located in that area of the continuum where the acrolect shades into the mesolect (cf. section 3.1 above and Deuber 2009b for Jamaica). In the quantitative analyses of ICE-T&T data presented here, class lessons were shown to be close to the conversations in language use, while little or no use of the Creole variants of the variables in question was found in the two monologic text categories from ICE-T&T included in the analyses, unscripted speeches and broadcast news. Thus, a quantitative approach in the variationist framework is certainly useful to locate a whole set of data within the continuum. It can also help to assess the level of language use in individual texts (cf. section 3.3) but many grammatical variables are not frequent enough to be analysed in each text in a category separately. However, another aspect of analysis that the quantitative approach is useful for is to compare different features to determine their relative 'Creoleness'. This then provides the background against which language use in individual texts can be interpreted. The more marked Creole features are most likely to have a special function in discourse. A qualitative approach shows that there are important differences in the discourse functions of Creole features in different categories of texts. They are most varied in the conversations, and in this text category especially it can be fruitful to apply an approach which considers identity construction in discourse as an explanatory factor for speakers' choices. Quantitative and qualitative

approaches thus complement each other in the analysis of Creole features in Caribbean ICE corpora.

Mair (2007) has proposed a greater integration of corpus linguistic and sociolinguistic approaches to language variation, pointing out affinities between recent work in corpus linguistics and the quantitative variationist paradigm in sociolinguistics and illustrating these with selected analyses of ICE-Jamaica data. The present paper has taken the integration of corpus linguistics and sociolinguistics one step further by showing that not only a quantitative sociolinguistic approach but also a qualitative one which focuses on the discourse functions of relevant features can be profitably applied to data from Caribbean ICE corpora. Such an interpretive approach may sometimes require more background knowledge than corpus linguists, who are not members of the speech community that their corpus represents, may have, but one of the best ways of at least partially overcoming this obstacle is to combine fieldwork – usually the domain of the sociolinguist – and corpus linguistics by getting involved in the compilation of one's corpus.

Notes

1 The fieldwork and analyses reported on in this paper were supported by a fellowship from the German Academic Exchange Service (DAAD) to the author for a one-year stay at the University of the West Indies in St. Augustine, Trinidad (where a most propitious working environment was provided by the Department of Liberal Arts, then headed by Prof. Valerie Youssef), by a grant from the Campus Research and Publication Fund, University of the West Indies, St. Augustine, to Valerie Youssef and the author for work on the Trinidad and Tobago component of ICE, and by a grant from the *Deutsche Forschungsgemeinschaft* to Christian Mair for the research group "Educated Spoken English in Jamaica" (DFG MA 1652/4).

2 Readers interested in the composition of the written component are referred to Nelson (1996) or www.ucl.ac.uk/english-usage/ice/design.htm.

3 ICE-Jamaica, which is being compiled at the English Department, University of Freiburg, in collaboration with the Department of Language, Linguistics and Philosophy, University of the West Indies, Mona, is close to completion. ICE-T&T, a joint project of the Department of English, University of Freiburg, and the Department of Liberal Arts, University of the West Indies, St. Augustine, was launched in 2006.

4 This is in contrast to written usage, where the Creole plays only a minor role, as Mair (2002) has shown.

5 Winford (1980: 56) points out the fact that social class I has higher values than social class II and explains this by an exceptionally high incidence of unmarked forms in the speech of one informant in class I.

6 Cf. also Labov (2006: chapter 4).

7 In the category 'broadcast news' only speech by news presenters and reporters (ca. 30,000 words) has been analysed. Speech by other persons often represents different text types, e.g. when extracts from parliamentary debates are included in the news.

8 The analysis is the same as applied by Deuber (2009b) to ICE-Jamaica data.

9 The analysis is limited to present-tense forms since in past contexts Trinidadian Creole regularly has invariant *was* as a copula form (cf. Winford 1992: 50).

10 Fifteen of the class lessons included in ICE-T&T (S1B-001 to 015) were recorded in secondary schools while the remaining five (S1B-016 to 020) are from the University. Since ICE aims to represent the language of adults who have completed secondary education or higher (cf. Greenbaum 1996: 6), students' speech in S1B-001 to 015 was marked as extra-corpus speech and excluded from the analyses, but accommodation to the students on the part of the teachers is, of course, a factor to be considered (cf. also Deuber 2009a).

11 James and Youssef (2004: 466) do not include these forms at all in their table of pronoun forms in 'mesolectal Trinbagonian', though they are used in the basilect, which in T&T is restricted to Tobago (cf. James and Youssef 2004: 466). Solomon (1993: 48) describes the use of *me* as subject pronoun as "severely restricted", possibly to a single idiomatic expression.

12 In the case of ICE-Jamaica there is a certain overlap between these hierarchies as one of the overt forms (*them* as subject pronoun) is actually more frequent than one of the zero forms (zero copula with adjectival predicates). For a fuller discussion of the hierarchy of Creolisms in the ICE-Jamaica data, see Deuber (2009b).

13 The change in speaker identification is due to the fact that in ICE letters are assigned to the speakers according to their order of appearance, starting anew for each text.

14 Short for Caribbean Institute of Media and Communication.

15 The speaker whose speech is marked as extra-corpus speech (<X> ... </X>) is actually a Trinidadian, but he had spent several years abroad.

16 The relative frequency is close to zero, therefore it is not apparent in figure 5 that this form occurs in the T&T data at all.

References

Allsopp, R. (1996), *Dictionary of Caribbean English Usage.* Oxford: Oxford University Press.

Coupland, N. (2007), *Style: Language Variation and Identity.* Cambridge: Cambridge University Press.

Deuber, D. (2009a), 'Standard English in the secondary school in Trinidad: problems – properties – prospects', in: T. Hoffman and L. Siebers (eds.) *World Englishes: Problems – Properties – Prospects.* Amsterdam: John Benjamins.

Deuber, D. (2009b), '"The English we speaking": morphological and syntactic variation in educated Jamaican speech', *Journal of Pidgin and Creole Languages*, 24: 1-52.

Deuber, D. and V. Youssef (2007), 'Teacher language in Trinidad: a pilot corpus study of direct and indirect Creolisms in the verb phrase', in: *Proceedings from the Corpus Linguistics 2007 Conference.* www.corpus.bham.ac.uk/corplingproceedings07/.

Görlach, M. (1991), 'English as a world language – the state of the art', in: M. Görlach (ed.) *Englishes: Studies in Varieties of English 1984-1988.* Amsterdam: John Benjamins. 10-35.

Greenbaum, S. (1996), 'Introducing ICE', in: S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon. 3-12.

Holmes, J. (1996), 'The New Zealand spoken component of ICE: some methodological challenges', in: S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon. 163-181.

Hudson-Ettle, D.M. and J. Schmied (1999), *Manual to Accompany the East African Component of the International Corpus of English.* www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/index.html.

James, W. and V. Youssef (2004), 'The Creoles of Trinidad and Tobago: morphology and syntax', in: B. Kortmann and E.W. Schneider (eds.) *A Handbook of Varieties of English*, *Volume 2: Morphology and Syntax*. Berlin: Mouton de Gruyter. 454-481.

Labov, W. (2006), *The Social Stratification of English in New York City.* 2nd ed. Cambridge: Cambridge University Press.

Mair, C. (1992), 'Problems in the compilation of a corpus of Standard Caribbean English: a pilot study', in: G. Leitner (ed.) *New Directions in English Language Corpora: Methodology, Results, Software Developments.* Berlin: Mouton de Gruyter. 75-96.

Mair, C. (2002), 'Creolisms in an emerging standard: written English in Jamaica', *English World-wide,* 23: 31-58.

Mair, C. (2007), 'Corpus linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change'. Paper presented at *ICAME 28*, 23-27 May, Stratford-upon-Avon.

Nelson, G. (1996), 'The design of the corpus', in: S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon. 27-53.

Nelson, G. (2002), *Markup Manual for Spoken Te*xts. www.ucl.ac.uk/english-usage/ice/manuals.htm.

Patrick, P.L. (1999), *Urban Jamaican Creole: Variation in the Mesolect.* Amsterdam: John Benjamins.

Schmied, J. (1996), 'Second-language corpora', in: S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon. 182-196.

Solomon, D. (1993), *The Speech of Trinidad: A Reference Grammar.* St. Augustine: University of the West Indies School of Continuing Studies.

Winford, D. (1972), *A Sociolinguistic Description of Two Communities in Trinidad*. PhD thesis, University of York.

Winford, D. (1980), 'The Creole situation in the context of sociolinguistic studies', in: R.R. Day (ed.) *Issues in English Creoles: Papers from the 1975 Hawaii Conference.* Heidelberg: Groos. 51-76.

Winford, D. (1992), 'Another look at the copula in Black English and Caribbean Creoles', *American Speech*, 67: 21-60.

Youssef, V. (2002), 'The language range of the Tobagonian people', in: W. James and V. Youssef (eds.) *The Languages of Tobago: Genesis, Structure and Perspectives.* St. Augustine: University of the West Indies School of Continuing Studies. 12-39.

Youssef, V. and D. Deuber (2007), 'ICE Trinidad and Tobago: teacher language investigation in a university research class', in: *Proceedings from the Corpus Linguistics 2007 Conference*. www.corpus.bham.ac.uk/corplingproceedings07/.

Appendix

Table 1a: Selected morphological and syntactic features of Jamaican and Trinidadian Creole

	Jamaican Creole	Trini-dadian Creole	Example (from the data cited in the text)
Morphology/ verb phrase :			
-progressive	-ing, a	-ing	we liming (9)
-habitual	Ø	Ø, does	I does talk me Creole (9)
-past	Ø, did, ben	Ø, did	say (6); did try (7)
-completive	done	done	you done trap in your ways (8)
-future	wi, go	go	if they go tell me (8)
-3rd sing. present	Ø	Ø	the situation warrant it (9)
-negation	no (and other forms)	ain't (and other forms)	me no understand (5); he ain't know (8)
-copula_adj, present	Ø	Ø	they aware (8)
-copula_locative, present	Ø, de	Ø, (de)	you in front (9)
Morphology/ noun phrase:			
-plural	them	them, and them	children and them (9)
-1st sing. subj. pronoun	me	I	me no understand (5)
-2nd pl. subj. pronoun	una	all you	all you dissing me (4)
-3rd pl. subj. pronoun	them	they, (them)	them lock up themself (8)
-1st sing. poss. pronoun	me	me	me time (3)
Syntax:			
-questions	no do-support/ inversion	no do-support/ inversion	what that has to do (7); why you're making (7)

Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora[1]

Alpo Honkapohja, Samuli Kaislaniemi and Ville Marttila

University of Helsinki

Abstract

This paper introduces a new project, Digital Editions for Corpus Linguistics (DECL), which aims to create a framework for producing online editions of historical manuscripts suitable for both corpus linguistic and historical research. Up to now, few digital editions of historical texts have been designed with corpus linguistics in mind. Equally, few historical corpora have been compiled from original manuscripts. By combining the approaches of manuscript studies and corpus linguistics, DECL seeks to enable editors of historical manuscripts to create editions which also constitute corpora.

The DECL framework will consist of encoding guidelines compliant with the TEI XML standard, together with tools based on existing open source models and software projects. DECL editions will contain diplomatic transcriptions of the manuscripts, into which linguistic, palaeographic and codicological features will be encoded. Additional layers of contextual, codicological and linguistic annotation can be added freely to the editions using standoff XML tagging.

The paper first introduces the theoretical and research-ideological background of the DECL project, and then proceeds to discuss some of the limitations and problems of traditional digital editions and historical corpora. The solutions to these problems offered by DECL are then introduced, with reference to other projects offering similar solutions. Finally, the goals of the project are placed in the wider context of current trends in digital editing and corpus compilation.

1. Introduction

The Digital Editions for Corpus Linguistics (DECL) project aims to create a framework for producing online editions of historical manuscripts suitable for both corpus linguistic and historical research. This framework, consisting of a set of guidelines and associated tools, is designed especially for small projects or individual scholars.

A completed DECL edition will, in effect, constitute a lightly annotated corpus text. In addition to a faithful graphemic transcription of the text itself, DECL editions will also contain information about the underlying manuscript reality, including features like layout and scribal annotation, together with a nor-

malised version of the text. All of these features, encoded in standoff XML, can be used or ignored while searching or displaying the text.

DECL was created in 2007 by the three authors, who are postgraduate students at the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki. We shared a dissatisfaction with extant tools and resources, believing that digitised versions of historical texts and manuscripts generally failed to live up to expectations. At the same time, we recognised that digitisation was time-consuming and complicated, and thus compromises had been made in the creation of digital editions and corpora. In order to alleviate these problems, we began the design of a user-friendly framework for the creation of linguistically oriented digital editions using extant standards, tools and solutions.

The first three DECL editions will form the bases for the doctoral dissertations of the writers. Each of these editions – a Late Medieval bilingual medical handbook (Honkapohja), a family of fifteenth-century culinary recipe collections (Marttila), and a collection of early seventeenth-century spy letters (Kaislaniemi) – will serve both as a template for the encoding guidelines for that particular text type and as a development platform for the common toolset. The editions, along with a working toolset and guidelines, are scheduled to be available within the next five years.

This article attempts to outline some of the problems and limitations of traditional digital editions and historical corpora. The core problem is the lack of integration between the disciplines of manuscript editing and corpus linguistics, which results in duplication of effort and loss of linguistically valuable data. The DECL project is intended to help eliminate the rift between these two disciplines by promoting editions of historical texts that take into account the requirements of corpus linguistics, and corpora that accurately represent the reality of historical documents. In practice this means defining a framework for producing detailed digital editions which not only reproduce all aspects of the original document but can be compiled into corpora without need for further processing.

2. Theoretical and ideological orientations

Editing involves making decisions which are practical on the surface, but have underlying hermeneutic and theoretical implications (cf. e.g. Machan 1994: 2-5). When the aim is to create digital editions which encode a wide range of manuscript-related phenomena into standardised XML mark-up, the challenge to editorial principles is significant. The issue is further complicated by the heterogeneous target audience: historians and linguists can have widely differing assumptions about what constitutes data and how it should be presented. Consequently, it is necessary to outline the underlying theoretical orientations of the DECL project, and to place them in the context of theory and bibliographical practice within the field.

2.1 Artefact, text and context

In order to conceptualise and model the various types of information encoded into a DECL edition, we use a three-fold division of 'artefact', 'text' and 'context'. With 'artefact' we are referring to the actual physical manuscript, by 'text' to the linguistic contents of the artefact, and by 'context' to both the historical and linguistic circumstances relating to the text and the artefact. This division originates in the discussion of a similar categorisation in Shillingsburg (1986: 44-55), and especially in its practical application by Machan (1994: 6-7), when editing Middle English texts. (By using 'artefact' for what Shillingsburg terms 'document', we aim to avoid confusion and overlap with the widely accepted meaning of 'document' in linguistic computing: the electronic text created by the editor). The concepts of text and artefact also roughly coincide with the terms 'expression' and 'item' defined in *Functional Requirements for Bibliographic Records* (FRBR: 13).

Since DECL is concerned with what Shillingsburg calls the historical orientation of editing (1986: 19), our starting point and primary focus is the individual artefact. We see the text as a cultural product and interesting in itself, not merely as a manifestation of a work of art produced by an individual author, on which systems like FRBR tend to focus. The concept of 'a work' is not a simple one, and may create more problems than it solves when dealing with texts like personal letters or a collection of anonymous culinary recipes written down in several hands. The question of authorship can also be problematic with medieval and Early Modern texts. As a result, we have decided to omit both categories, since they run the risk of making the framework too rigid to deal with non-literary historical manuscript texts.

On the other hand, our focus on texts as cultural products has led us to add the concept of 'context' to represent the outside circumstances related to the production and use of the artefact and the text. Context covers the various types of cultural, social and historical background material and bibliographical information that is included in an edition. Our terms are designed to illustrate the interrelationships of the different types of features that are encoded in a DECL edition. They are meant as fuzzy rather than rigid categories, and serve to theorise how the non-linguistic aspects – including historical, codicological and bibliographical – relate to the textual whole. They serve as the foundation for a model of editing that aims to be comprehensive enough to cover all of the tasks involved in editing historical manuscripts, yet flexible enough to be adaptable to the needs of different editing projects.

2.2 Editorial principles

The field of historical linguistics has seen some recent discussion over what is required of an edition or corpus for it to be suitable for historical linguistic study (cf. e.g. Bailey 2004; Curzan and Palmer 2006; Dollinger 2004; Grund 2006). Most vocal in his criticism of existing practices is Lass (2004), who demands

that, in order to serve as valid data for the historiography of language, a digital edition or a corpus should not contain any editorial intervention that results in substituting the scribal text with a modern equivalent. He gives examples of several commonplace editorial practices, such as invisible emendations, silent expansion of abbreviations, modernisation of punctuation and word division, and attempts to construct lost archetypes based on multiple manuscript witnesses. All of these deny the reader access to information present in the manuscript original, and instead create a new artificial language variant (Lass 2004: 22). To avoid this, Lass (2004: 40) defines three criteria which he considers inviolable for a historical corpus:

- maximal information preservation;
- no irreversible editorial interference;
- maximal flexibility.

While being very polemic, Lass does raise useful points and expose a number of harmful practices within historical linguistic study. It is clear that the requirements he proposes are something that compilers of editions should take into account, and therefore we have used them as a starting point, developing them further into three principles: 'flexibility', 'expandability' and 'transparency'. In actual practice, these three principles influence practical considerations such as tagging, data structure, and interface design.

- Flexibility: DECL editions seek to offer a flexible and user-friendly interface, which will allow the user to select the features of the text, artefact and context to be viewed or analysed. All editions produced within the DECL framework will build on similar logic and general principles, which will be flexible enough to accommodate the specific needs of any text type.
- Transparency: The user interface of DECL editions will include all the features that have come to be expected in digital editions. But, in addition to the edited texts and facsimile images of the manuscripts, the user will also be able to access the base transcripts and all layers of annotation. This makes all editorial intervention transparent and reversible, and enables the user to evaluate any editorial decisions. In addition, the DECL framework itself will be extensively and clearly documented.
- Expandability: DECL editions will be built with future expansion and updating in mind. This expandability will be three-dimensional in the sense that new editions can be added and linked to existing ones, and both new documents and new annotation layers can be added to existing editions. Furthermore, DECL editions will not be hardwired to a particular software solution, and their texts can be freely downloaded and processed for analysis with external software tools. The editions will be maintained on a web server and will be compatible with all standards-compliant web browsers.

The DECL framework, however, goes further than Lass (2004), which is primarily concerned with retaining the original scribal text. We adopt a similar position also with respect to the artefact and its context, which are treated as equally im-

portant aspects of manuscript reality and subject to the same qualitative requirements as the text itself.[2]

3. Limitations of earlier digital editions

An increasing number of digital editions of historical texts is being published online, which, for the linguist, is a mixed blessing. On the one hand, access to a greater number of texts on the web obviously makes new areas of research possible, on the other, not all of the editions are amenable to linguistic enquiry. What a linguist would ideally need includes:

- access to the language of the original text in unadulterated form;
- full text searchability with sortable and refinable search results; and
- the possibility of defining and, preferably, extracting sub-corpora.

However, digital editions range from those providing only facsimile images to those which have interfaces with most, if not all, of the features listed above.

Facsimile editions are a type of digital edition common, in particular, to repositories – namely libraries and archives – which are primarily concerned with the preservation and sustainability of digitised resources. Examples include the *Papers of Joseph Banks* at the State Library of New South Wales, but also the *Boyle Papers Online* at Birkbeck College, University of London. Editions including both facsimile images and transcripts of all or some of the texts, such as the *Auchinleck Manuscript* or the *Hooke Folio Online*, are more useful to linguists as they usually have (often limited) search functions, but these editions rarely allow users to download all the texts. Examples of editions with more elegant interfaces include the *London Provisioner's Chronicle* (the diary of Henry Machyn) and the *Letters of Clemency from the Chancery of Brittany*. This last example is the best from a linguist's viewpoint, for it not only provides the online user with facsimile images, diplomatic transcripts and indexes, but also allows the user to download the entire edition.

Commercial publications tend to have better functionality than freely available online editions. Examples of these for single works are the *Canterbury Tales* and other editions produced by Scholarly Digital Editions and Evellum, and for larger projects, *State Papers Online*.[3]

The great variety among digital editions is somewhat alleviated by the fact that, unlike most historical corpora (on which see section 4.7 below), many do use TEI XML – but few make full use of the potential of their XML encoding.[4] What the DECL framework aims to do is to increase comparability across the board by encouraging editors to cater to the 'linguist's needs' listed above. The framework is primarily intended for small-scale projects, which usually lack extensive funding, and which would greatly benefit from the development of more user-friendly tools and guidelines (cf. Robinson 2005).

4. Problems with traditional historical corpora

Much of the inspiration for the DECL framework comes from the shortcomings of traditional historical corpora as perceived from the point of view of a textual scholar. Most of the problems associated with using traditional historical corpora stem from the fact that because the transcription and digitisation of original manuscript texts into machine-readable form takes a lot of time and expertise, most historical corpora are based on printed editions, which "have generally not been produced with linguistic study in mind, and may not always be reliable" (Kytö et al. 2007: section 3).

4.1 Use of critical editions

The most obvious problem occurs when corpora are based on critical editions which compound multiple manuscript witnesses into a single text. Compiling a corpus from these types of editions multiplies the problems inherent in them. Combining elements from several textual variants, and potentially widely differing dialectal features or scribal practices, introduces a layer of linguistic hybridity which represents the language of the editors, not of the text.

In the best case scenario, the limitations are clearly documented and acknowledged; in a more likely case, the inclusion of the text into the corpus obscures the textual nature of the edition used and thus also the unsuitability of the text for many linguistic research questions. Despite the prevalence of critical editions in the tradition of textual editing, some text types – such as letters and other unique documents – form an exception by being frequently available as single-witness editions, thus avoiding the problem of linguistic hybridity and being suitable for the study of at least morphology and syntax, as well as pragmatics (Nurmi 1999: 55).

4.2 Varying editorial principles and loss of manuscript features

Another problem related to the use of editions is caused by their varying editorial principles. This problem is especially acute in corpora containing texts both from editions of varying types and from original manuscript sources. A prime example, despite all its strengths, is the important and pioneering *Helsinki Corpus of English Texts* (HC). As Kytö (1996: section 2) points out, "editorial and typographical conventions vary in different source texts (e.g. emendations can be indicated by italics, parentheses, brackets etc.)", and "a number of 'text level' codes have been used to transfer the function of the convention to the computerised version, irrespective of the particular format followed in the source text".

Although this kind of practice would, on the surface, seem to produce a uniform result, the format used, the amount of editorial intervention, and the degree to which various features of the original manuscript have been included can vary significantly between texts. Not many printed editions of prose texts reproduce the original layout of the text even to the level of manuscript lineation, and even fewer indicate textual details such as hand changes, scribal emendations or

abbreviations. This will result in either corpus texts having a variable amount of detail encoded in them, or omitting detail from those texts that would have it. The latter phenomenon is visible in HC, which has omitted original folio and page changes, customary for most critical editions.

In either case, textual or physical features that are not recorded in all of the editions cannot be used for analysis. The worst case scenario in this respect would be a corpus that encodes the features found in each edition without any information about the principles behind the editorial decisions. Fortunately, many corpus compilers do recognise the heterogeneous nature of the corpus contents:

> Editing policies vary a great deal during the 160-year history of editing medical texts, the scope being from the construction of hypothetical "originals" to faithful transcriptions. MEMT represents the "edited truth" of the underlying manuscript reality and we have reproduced the editions according to our principles […]. Thus the texts are twice removed from their manuscript reality. (*Middle English Medical Texts* (MEMT): Introduction)

4.3 Predetermined research focus

Reliance on existing editions, regardless of their editorial principles, results in another type of problem, often overlooked perhaps because of its obvious nature, namely that a corpus based on edited texts is, by necessity, circumscribed in its material by what has previously been considered worth editing. Textual editors tend to focus on texts considered culturally or literarily 'significant', and relying solely on editions can lead to the omission of whole categories of material. As the compilers of both the *Corpus of Early English Correspondence* (CEEC) and MEMT note, this problem is not limited to the realm of literary texts but affects all genres of historical writing:

> A more unexpected problem is the penchant, particularly of 19th-century editors, to edit only the letters of historically important people, and ones describing important historical events. Editors often disregarded family letters concerning everyday life, which would serve as better material for historical sociolinguistics. (Nurmi 1999: 54)

> Choices made by early editors tend to define the contents of e.g. literary and linguistic histories. In language histories, the early phases of scientific writing are often ignored or passed over with few comments for the simple reason that writings in this register were not known to researchers of the time. (MEMT: Introduction)

4.4 Questionable orthography

The use of printed editions presents several problems on the level of the text itself. The most obvious of these, prevalent especially in older editions, is the question of orthography. Few pre-1980s editions provide detailed information about their practices concerning orthography and frequently normalise spelling – not to mention punctuation – to varying degrees. While the regularisation of spelling may help with problems related to spelling variation and automated linguistic analysis, it also means that, as a rule, corpora based on printed editions cannot be used for the study of orthography or any other research question dependent on original spelling, as noted by the compilers of the *Corpus of Early English Correspondence Sampler* (CEECS):

> Particularly the older editions (ie the ones included in the CEECS) cannot be relied upon in questions of spelling, as the editors' priorities were often not linguistic but historical. Even […] newer editions […] [may be] a less than reliable source for studies of orthography. (Nurmi 1999: 55)

4.5 Copyright issues

In addition to the aforementioned problems relating to the integrity of the text, the use of printed editions also involves problems concerning the compilation and publication of corpora. Perhaps the most restricting of these is the problem of copyright. While historical documents (at least from the Medieval and Early Modern periods) are free of copyright, modern printed editions of these documents usually are not. This leaves the corpus compiler with two options: either use old, out-of-copyright editions or contact the publisher (or other copyright holder) of a more recent edition for permission to include the material in a corpus, often involving a considerable fee.[5]

Both of these approaches have their problems. Editions from the nineteenth or early twentieth century, which are now in the public domain, often fail to meet the standards required of reliable data for historians or historical linguists, and using them will exacerbate many of the problems mentioned above.

On the other hand, since the texts in traditional historical corpora often come from a variety of sources, obtaining permission from all copyright holders can be a daunting task. For instance, Kytö (1996: Preface) expressly acknowledges the generosity of 38 separate persons, publishers and institutions for providing permission to include their texts in the HC. Contacting copyright holders can be very difficult and time-consuming. The corpus compiler may encounter situations where the rights have moved from one holder to another or where the institution holding them has ceased to be operational, and, in the end, the current holder may or may not grant them (see e.g. MEMT: Introduction; Nurmi 1999: 56).

One partial solution or way of working around the problem of copyright is the distribution of tasks through collaboration when compiling corpora – an approach taken by the Salem Witchcraft Records corpus and the updated version of *A Representative Corpus of Historical English Registers* (ARCHER). While the resulting corpus cannot be freely distributed, this method increases the number of institutions where the corpus is available.

4.6 Duplication of effort

Two more problems that stem from using printed editions in compiling corpora are the duplication of effort and an increased probability of errors. Producing an edition of manuscript material in whatever form involves a significant amount of work. If the edition is published in printed form and used as a source for a corpus, the compiler will need either to key in the whole text or use Optical Character Recognition (OCR) software to digitise it. Both of these methods require at least some degree of proofreading and are likely to introduce new errors into the text. This kind of perceived waste of effort was actually one of the key issues in forming the DECL project: it was important to ensure that digital editions would be immediately useable as corpus texts without a significant amount of additional work.

4.7 Problematic corpus conventions

Traditionally, corpora have been viewed as monolithic entities – collections of texts that are compiled, digitised and annotated, and when all the stages are finished, released as a whole.[6] As a result, large or otherwise work-intensive corpora can spend years as 'work-in-progress', being generally unavailable to the scholarly community even if significant parts of them are already finished. Furthermore, this view of corpus compilation as a large undertaking involving a huge mass of texts can easily discourage small projects and individual scholars from compiling corpora, because it would take too long to compile a corpus of sufficient size.

Once a corpus is finished, it is not commonplace to make provision for including new material. There have been several updated or expanded versions of earlier corpora, but even they have mostly taken the form of new, individual, closed products.[7] DECL aims to provide means to add new content, either in the form of new texts ('horizontal expansion'), additional annotation ('vertical expansion') or supporting background material.

The requirements posed by this kind of expandability also reveal another problematic property of many corpora, namely the use of corpus- or project-specific tagging and encoding practices, often developed for the needs of one specific corpus. Although there are some accepted and established principles and ways of encoding corpus material, the situation in the case of corpora is far from the optimistic view that seems to prevail in the field of digital humanities: "gone, too, are the days when every individual or project invented codes, systems, or symbols of their own to identify special features, and any character that could not

be represented in ASCII had to be recoded in some arcane form" (Deegan and Tanner 2004: 493-494).

This seems to be mainly a historical development. Many corpora have borrowed their encoding and mark-up practices from earlier corpora and adapted them to their own use.[8] This kind of variance limits the development and use of common tools and the convertibility of corpora from one format to another. The situation is somewhat surprising, considering that standards for the electronic encoding of textual data, most notably the Text Encoding Initiative (TEI) have been around for almost two decades (the first version of the TEI Guidelines was published in 1990). There are some historical corpora that use a version of the TEI Guidelines, such as *The Lampeter Corpus of Early Modern English Tracts*, but use of the Guidelines seems to be significantly more common in other branches of digital humanities than in corpus linguistics.

4.8 Shallow representation of manuscript reality

Historical corpora are often characterised on a two-dimensional scale as 'long' or 'short', and 'thin' or 'fat', the first categorisation reflecting their diachronic scope and the second the extent of their synchronic coverage (cf. Rissanen 2000). Comparatively less attention has been paid to a third dimension, 'depth', which could be defined as the extent to which the corpus represents the various features of the original texts. This dimension is especially relevant in the case of materials with limited availability, such as historical manuscripts. A deeper representation helps to widen the applicability of the corpus to different types of research, which is important for specialised corpora that run the risk of becoming marginal if their applicability is further limited by design or compilation choices.

Moreover, in contrast to digital editions, most linguistic corpora have given little attention to the visual presentation of text, being oriented towards linguistic analysis. This, together with the limited search and analysis tools provided by most digital editions, has created a wide but unnecessary rift between these two types of digital resources, which at their heart have much in common and could both benefit immensely from closer integration with each other.

5. Key features of the DECL framework

As a response to these problems and driven by the theoretical and ideological orientations described above, the DECL framework has been designed to overcome the limitations and combine the benefits of both digital editions and traditional historical corpora. Most of the individual features described below are not unique to DECL but are evidenced by various other corpus and digital editing projects. The aim of the DECL framework is to learn from the example of these projects and to bring together their best aspects while simultaneously avoiding as many of the abovementioned problems as possible.

5.1 Faithful representation of original texts

Since the DECL framework is intended for producing digital editions of historical texts, one of its primary objectives must be the definition of clear and consistent editorial principles. Being oriented primarily (though not exclusively) towards producing editions useful for corpus linguistics, the emphasis must be on representing authentic language use. The need for more linguistically-oriented editions that "aim at reproducing the original manuscripts more faithfully than critical or eclectic editions do" (Kytö et al. 2007: section 3) has been widely acknowledged in recent years. This has also affected the compilation principles of many recent corpus projects, such as the *English Witness Depositions 1560-1760: An Electronic Text Edition* (EWD) project at the University of Uppsala, the *Middle English Grammar Corpus* (MEG-C) at the University of Stavanger, the *Linguistic Atlas of Early Middle English* (LAEME) and the *Linguistic Atlas of Older Scots* (LAOS) at the University of Edinburgh, the *Corpus of Early Ontario English* (CONTE) at the University of British Columbia, *A Corpus of Middle English Scientific Prose* (ACOMESP), a collaboration between the University of Málaga and the University of Glasgow, and the *Corpus of Scottish Correspondence* (CSC) at the University of Helsinki.[9]

The editors of EWD introduce the concept of a "linguistic edition" and define it as an edition where "the language of the original manuscript text is not normalised, modernised, or otherwise emended", but "the manuscript is reproduced as closely as possible in transcription" (Kytö et al. 2007: section 3). Similarly, the compilers of the MEG-C aim "to record what is visible in the manuscript, rather than giving editorial interpretations" (Stenroos and Mäkinen 2008: 14), reproducing the text "at what might be called a rich diplomatic level" (Stenroos and Mäkinen 2008: 7). This type of linguistic edition is essentially what lies also at the heart of a DECL edition: a diplomatic transcription of an individual manuscript witness, representing a sample of authentic language use.[10] In the case of DECL and both of the abovementioned projects, this also entails the use of original manuscripts as the source, although microfilms and digital reproductions can be used as an aid in the editing process.

Editions produced using the DECL framework will preserve the orthography of the original manuscript down to graphemic level without normalising either spelling or punctuation. The DECL guidelines also aim at the preservation of the original word-division, but since the word-spacing of manuscript texts is not always reproducible in digital format, editorial judgement of whether two words are separated by a space will be required in unclear cases. While preserving the original orthography, the DECL framework will also provide tools and guidelines for annotating every word token of the original text with its normalised form, facilitating searches and automated analysis of the text with tools developed for Present-day English.

Since the DECL framework places equal emphasis on the levels of text, artefact and context, the scope of faithful representation extends beyond the strictly textual level. DECL editions will try to represent the physical layout and

appearance of the text on the manuscript page – ideally both as machine-readable tagging and in facsimile images – and provide a description of the cultural and historical context of the text.

Another aspect of faithful representation, which has traditionally been associated with digital editions rather than corpora, is the visual representation of textual and palaeographical features of the text. DECL editions will have an online interface which will be customisable in two senses. Firstly, the editors of individual DECL editions will be able to choose which features will be implemented in their edition and, since all tools developed for the DECL framework will be open source, even program new features. Secondly, the interface will enable the user to choose the features of the text to be viewed, downloaded or included in the analysis. In addition to visual presentation and browsing, the interface will also offer corpus search and analysis functions and the ability to download the texts in various formats.

5.2 Edition = corpus text

As pointed out above, one of the central ideas behind the DECL project is to combine the strengths of digital editions and linguistic corpora into a single multi-purpose resource. Considering that many digital editions and all historical corpora are essentially digital transcripts of text, whose production involves quite similar tasks, there have been surprisingly few attempts to combine them. While it is true that many digital editions have rudimentary search tools and some corpora provide ways of visually representing the corpus texts, only a few projects attempt to create editions that would serve as corpora straight out of the box.

There are some important predecessors: two examples of projects with similar aims are the EWD and ACOMESP already mentioned above. The editors of the EWD emphasise that their edition "will be geared to facilitate advanced computer searches" and that they "combine [their] philological and editorial aims with principles of modern corpus compilation, striving at a new type of text edition that will also serve as a computerised corpus" (Kytö et al. 2007: section 5). ACOMESP in turn offers a web interface that allows facsimiles and transcriptions to be viewed side by side, as well as corpus searches to be conducted on the texts. The project benefits from being able to use high quality facsimiles from the Hunter collection of the library of the University of Glasgow.

What, then, are the basic requirements – in addition to the faithful representation discussed above – of an edition so that it can be used as (part of) a corpus? The most obvious requirement is for it to include machine-readable, i.e. digital, transcripts of the source texts. Next, it must be possible to perform text searches on the texts, preferably with support for regular expressions (or at least wildcards). This second requirement can be fulfilled either by including a suitable search engine in the interface or by allowing the text of the edition to be extracted in a format that is usable by external corpus tools – or, ideally, by both methods.

The elimination of the rift between an edition and a corpus also means that all of the textual and codicological features encoded in a DECL edition are automatically available in a corpus compiled of such editions without the need for

further encoding. This enables the corpus compiler to make full use of the work of DECL editors and to effectively leverage his or her expertise by focusing on the analysis and linguistic annotation of the data instead of its collection. Furthermore, all linguistic metadata added by the corpus compiler can also be made available for users of the original edition.

5.3 Modular and layered architecture

Since it is aimed especially at a community of small projects and individual scholars, the DECL framework promotes a view of corpora not as monolithic and closed text collections but as modular and flexible networks of texts, whose production can thus be distributed both in time and place.[11] In practice this means that by following the guidelines and practices defined by the DECL framework, independent scholars or projects can produce and release 'mini-corpora' or even individual texts, which can then be joined together into larger corpora and further supplemented with new texts. A similar process-like approach to corpus compilation has been adopted by the MEG project, although within a more traditional 'version paradigm' where each extended version of the corpus is seen as an individual product replacing its predecessor.[12] Releasing the corpus before it is 'finished' not only allows the scholarly community to benefit immediately from what has been accomplished so far, but also avoids limiting the potential size of the corpus: theoretically, new texts could be added until all known texts have been included.

The DECL guidelines have also been designed to allow for the addition of new layers of annotation to existing texts. This is made possible by the use of standoff annotation, where the annotation layers are maintained separate from the base text and linked to it by means of uniquely identified word tokens. These annotation layers are not limited to traditional linguistic annotation, but can contain any kind of ancillary information relating to the text.

This means that all editorial intervention and interpretation is not only indicated by mark-up, but also physically separated from the base text, rendering it transparent and easily reversible. While the use of annotation layers allows the user to focus on only the selected aspects of the text, they are also persistently linked together and can be freely accessed at any time. By allowing for the addition of new annotation layers to the text without changing the base text, the layered architecture not only ensures the stability of the base text, but also allows for the creation of mutually exclusive annotation layers.

In terms of corpus compilation, this means that once the number of DECL editions increases, corpora can be compiled simply by defining an annotation layer linking a set of DECL-compliant texts together. With equal ease, the corpus compiler can attach descriptive or classificatory attributes to individual texts, creating sub-corpora for comparative analysis.

To facilitate the automatic linguistic analysis of DECL editions, the framework calls for the inclusion of an annotation layer containing normalised forms for every word token, eliminating – or at least alleviating – the problem of spelling variation inherent in historical corpora.[13] Furthermore, the texts included

in the corpus can be analysed using external annotation tools, temporarily ignoring any annotation layers not relevant to the analysis. The results of this analysis can then be detached from the text and converted into a new annotation layer to be shared with others.

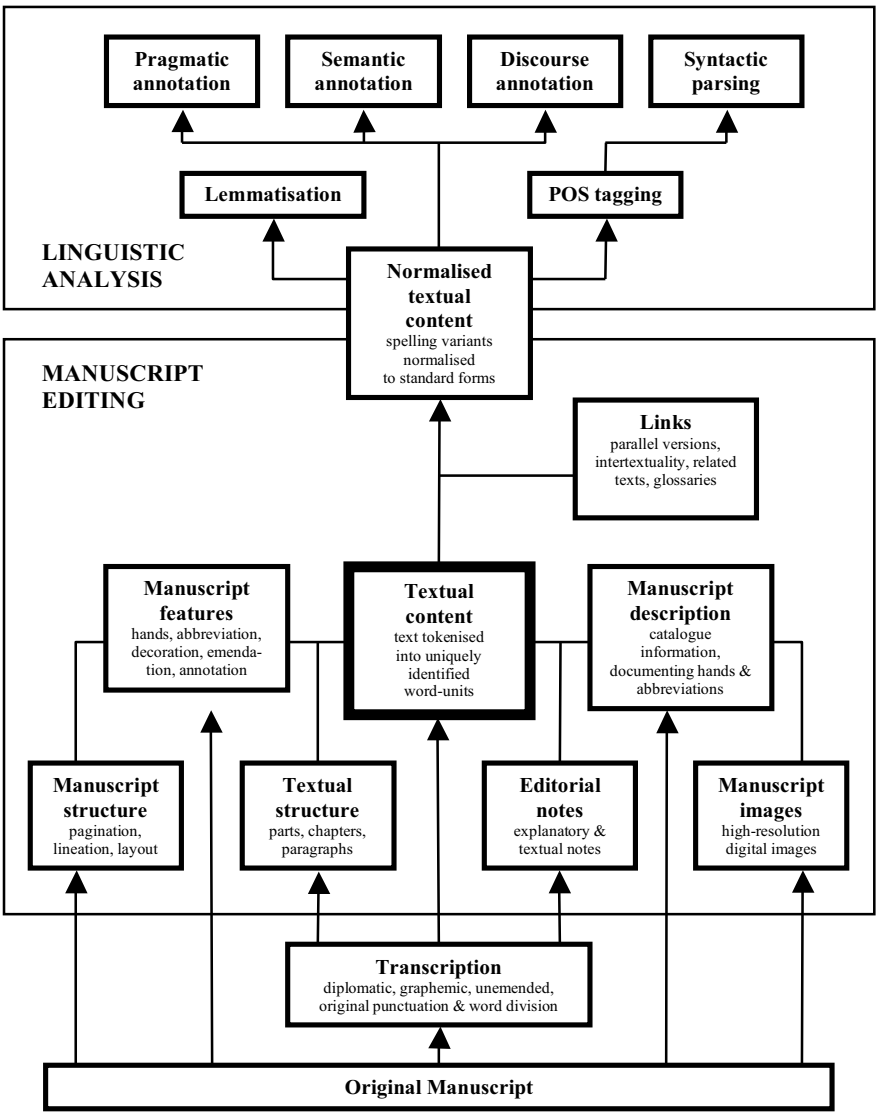

Figure 1: The conceptual structure of a DECL edition

This means that the user of a DECL edition – such as a corpus compiler – can extract the normalised version of the text, run it through a parser, and then link the parsed text into the edition as an additional annotation layer. This will enable future users to run syntactic queries on the parsed text, but have the results displayed in the original orthography and linked to their original manuscript context.

Figure 1 above illustrates the structure of a richly annotated DECL edition that has been included in a corpus and analysed for various linguistic features, along with the division of labour between the manuscript editor and the corpus linguist.

5.4 The virtues of standards

As pointed out earlier, the field of digital humanities has seen much work in the creation of encoding and mark-up standards in the last decade or so. Perhaps the most significant effort in providing standard forms of textual mark-up has been the TEI. Currently in their fifth public version (P5), the XML-based TEI Guidelines have been adopted by a large number of projects within the field of digital humanities, including the *British National Corpus* (BNC) and even some historical corpus projects, such as the *Corpus of Northern English Texts from Old to Early Modern English* at the University of Seville.[14] Expressed as a modular XML schema, the TEI Guidelines define a mark-up language for representing the structural, visual and conceptual features of texts.

The DECL framework is based on the TEI Guidelines, and the DECL editorial guidelines will be a strictly defined subset of the TEI schema, documented in detail. This means that any edition produced according to the DECL guidelines is automatically TEI-conformant and thus compatible with any TEI-compatible tools. Since the TEI and thus the DECL guidelines are valid XML definitions, more generic XML tools can also be readily used with DECL editions. Conversely, any tools produced within the DECL framework will also be usable (or can be modified to be so) within other TEI-compatible projects.

From a more technical viewpoint, XML brings several benefits. First of all, XML readily supports the kind of modular approach described above, and makes a clear distinction between the textual content and the mark-up, consisting of a defined set of elements, which can be described by assigning values to their attributes. Furthermore, XML mark-up provides the added advantage of using XSLT (eXtensible Stylesheet Language: Transformations) to manipulate and transform the content of documents, either to create new XML documents from the contents of the edition or to convert it into other mark-up formats. This will enable DECL editions to be used with various existing annotation, analysis and presentation tools. Furthermore, the XML mark-up used by the DECL framework does not restrict the annotator to any given linguistic annotation scheme, but can be used to encode a variety of schemes, such as CLAWS, CSC, NUPOS or Penn Treebank.

This makes the approach of DECL subtly but fundamentally different from that taken by other related projects, such as MEG-C (Stenroos and Mäkinen 2008: 15) or EWD (Kytö et al. 2007: section 2). Instead of releasing separate ver-

sions for different purposes (e.g. reading and linguistic searches), custom representations are created dynamically from the base XML according to the user's selections. This helps to maintain the link between all representations and the original data, meaning that, for example, any search results found using normalised forms of the words remain linked not only to the original forms but to all of the formatting and background information pertaining to them.

While the XML definition and the TEI Guidelines have largely standardised the technical aspects of encoding text, the aim of the DECL guidelines is to go further and to use these standards as a basis for defining and documenting a set of editorial principles and practices. This will eliminate the problem of varying editorial principles, discussed above, and allow texts edited following the DECL guidelines to be used together and combined into corpora.

5.5 Fundamental freedom

Since the DECL project is committed to the principles of open access and open source software, all of the tools and documentation of the DECL framework will be released following these principles as far as possible.[15] The project intends to both make use of existing open source software projects and adapt them to its needs, and develop new custom solutions for those needs that have not yet been met by existing solutions. All tools will be developed to be platform-independent and as flexible as possible.

Naturally, these principles will also be extended to any editions produced using the framework. Using original manuscripts as sources provides DECL editions freedom from external copyright: the copyright of the transcript resides with the transcriber. In order to avoid copyright issues between transcribers, editors and corpus compilers, and to allow DECL editions to be freely used in corpora, the framework requires that DECL editions be published under a suitable open access license. A similar approach has been taken for MEG-C, which is distributed under the Creative Commons Attribution-Noncommercial-Share Alike (*by-nc-sa*) license, giving the users freedom to not only use the corpus as it is, but also to create and publish derivative works under the same license, provided that the original work is credited to its authors and the derivative work is not distributed commercially.[16] This particular license is also the strongest candidate under consideration for publishing DECL editions.

This freedom extends also to the internal workings of the edition: in keeping with the idea of transparency, all layers of the edition from the base transcript to the various levels of annotation will be accessible for viewing, searching and downloading. This will not only ensure the reusability of previously created resources, but also enable the user to evaluate any editorial decisions.

Although using open access transcriptions of original sources solves the problem of copyright for the texts, the copyright of manuscript images remains a problem. Since most manuscript repositories[17] reserve the right to produce digital reproductions of their collections and charge significant fees for these reproductions, small projects, in particular, may be hard-pressed to obtain digital facsimiles even for their own use. Furthermore, since the repository that produced the

reproductions owns the copyright for them, they cannot be freely published under an open access license. The only way to get around this problem is to work with repositories and persuade them to either digitise the manuscript material and to publish them under an open access license, or to allow scholars to photograph manuscript material themselves.

With regard to corpus compilers, the DECL framework seeks to liberate them from the chains of 'what has been edited' and enable them to add texts from original sources with reasonable effort, effectively becoming digital editors themselves. It is clear that the viability of this depends on both the nature of the material and the text-scholarly competence of the scholars concerned. Yet while the DECL framework can offer only limited assistance in the textual scholarship required for editing original manuscript texts, it will provide a thoroughly documented mark-up for recording the features of the manuscript text, detailed guidelines on the various steps involved in creating a digital edition, and tools to facilitate and even automate many of the steps involved in turning a base transcript into a finished digital edition.

6. Conclusion: working towards mutual goals

We wrote above that DECL was triggered by a dissatisfaction with existing digital resources, and have argued that a more systematic effort should be made in the creation of digital resources of historical documents in order to increase their accessibility, usability and versatility. Similar concerns have been voiced by linguists and historians alike, as well as by archivists and other scholars. In the manual of the *Corpus of Scottish Correspondence* (CSC), Meurman-Solin writes that:

> [T]he fourth generation of corpora will combine three important properties. Firstly, we define language-external variables rigorously, benefiting from information provided by various interdisciplinary forums. Secondly, we see corpora as consisting of sub-corpora that are defined […] in reference to degrees of validity and relevance as regards their usefulness for the study of a specific research question. Thirdly, instead of marketing corpora as completed products, we see the compilation as an ongoing process, and therefore view expansion and revision as inherent characteristics of this work. (Meurman-Solin 2007, section 2.1.1)[18]

Meurman-Solin's second point is one pertinent to this age of web-based corpora used for studying Present-day English. Yet such an approach is becoming feasible for historical linguistics as well, as shown by De Smet's *Corpus of Late Modern English Texts* (CLMET), which he compiled from sources already available online:

> [T]he corpus can be extended or reduced at wish [sic], and similar – though not necessarily identical – corpora can be compiled without much effort by anyone […]. The corpus presented here is what I consider an acceptable and useful offshoot of a continual attempt to open up the rich resources of the Internet to historical linguistic research. (De Smet 2005: 70)

Still, CLMET is closer to a traditional historical corpus than CSC, in that its sources are digitised versions of editions of Late Modern English texts, while CSC is based on manuscripts. But, as mentioned above, one of the aims of the DECL project is to eventually enable the creation of historical corpora in a fashion similar to that of CLMET, based on a large number of DECL-compliant digital editions of historical documents. This objective is not a new idea, and has been dubbed a 'textbase approach' to using digitised resources (Vanhoutte and Van den Branden forthcoming).

In short, the aim is to make online resources into multi-functional databases by encouraging their creation according to defined standards. As Vanhoutte and Van den Branden (forthcoming: section 10) put it, "from a rich textbase of encoded […] material […] various derived products [can] be extracted and realised, such as scholarly editions, reading texts, indexes, catalogues, calenders, regests, polyfunctional research corpora etc". The textbase approach works in tangent with the concept of 'distributed' production: spreading the workload of a project by opening it to other scholars (as described above in section 5.3). Such collaboration would ultimately lead to shared online resources not entirely unlike Wikipedia (and other Wikimedia resources), but created and moderated by scholars for (primarily) scholarly purposes. These aims require collaboration at a high level, but fortunately such initiatives exist: one, for mark-up, is the aforementioned TEI; another, for general architecture, is the Distributed Editions Initiative led by the Institute for Textual Scholarship and Electronic Editing at the University of Birmingham.

The aims of DECL are much the same as those of the Distributed Editions Initiative: to create versatile digital resources by adhering to agreed standards, by allowing other scholars access to improve these resources, and by helping to create multidisciplinary shared online resources. In other words, we, too, are working towards "a federated model of scholarly tools and materials on the internet", as it is phrased on the Distributed Editions website (http://www.itsee.bham.ac.uk/DistributedEditions/summary.htm). While these theoretical goals may sound highly optimistic, in practice DECL hopes to participate primarily by creating more editions of previously unedited historical manuscripts, ensuring that all are suited for linguistic study.[19]

Notes

1 Work done on the DECL project has been funded by the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki, and by the Finnish Cultural Foundation.

2 Lass does acknowledge the importance of the features of the 'artefact' to some degree by mentioning the potential significance of retaining punctuation and manuscript page layout (2004: 36).

3 *State Papers Online* is arguably not an edition. However, while the features of its interface are similar to those of resources like *Early English Books Online* (EEBO) and *Eighteenth-Century Collections Online* (ECCO), its scope is more strictly defined. *State Papers Online* is also not to be confused with the admirable but lo-fi effort of the *State Papers Project*, a freely available edition of a part of the same material.

4 The *Letters of Clemency from the Chancery of Brittany* is an exception, being a simple but highly usable and versatile digital edition along the lines encouraged by the DECL project. Yet it shows particularly well what can be done with reasonable effort, and what functionalities all digital editions could have. The *Digital Archive of Letters in Flanders* (DALF) project editions produced by the Centre for Scholarly Editing and Document Studies (CTB) at Ghent also contain all these functionalities, yet are unfortunately not available online (Edward Vanhoutte: personal communication).

5 For examples of problems related to copyright in the context of corpus compilation, see e.g. *Middle English Medical Texts* (Introduction) and Nurmi (1999: 56).

6 The *Middle English Grammar Corpus* (MEG-C) is a welcome exception, as it is intended to be published in incremental parts as the work progresses. See section 5.3 and note 12.

7 For example, the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME) is built on the Middle English part of the HC, and has itself seen a second edition (PPCME2). There is also an updated version of ARCHER. An exception to the tendency of releasing updated corpora as complete entities is the *Corpus of Scottish Correspondence* (CSC), see section 6.

8 Examples of this include the *Middle English Medical Texts* (MEMT) corpus and the *Corpus of Early English Correspondence* (CEEC), which both use a mark-up system based on that of the *Helsinki Corpus of English Texts* (HC).

9 It is interesting – and to some degree indicative of the intimate relationship between historical corpus compilation and digital editing – that while the MEG-C and the *English Witness Depositions 1560-1760* are quite similar in their aims and methods, the former is described as a corpus and the latter as an electronic text edition.

10 This emphasis on individual manuscript witnesses does not preclude multi-text editions. The DECL framework will include provisions for producing editions of several manuscript versions of a text (or several closely related texts) and presenting them as a parallel text edition, enabling the comparison and analysis of the variation between versions.

11 Some of the difficulties involved in distributing the tasks of editing and corpus compilation between two completely separate projects without a common framework are exemplified by the interrelationship of the *Proceedings of the Old Bailey* and the *Old Bailey Corpus* as described by Huber (2007).

12 The first version of MEG-C, containing roughly a third of the base texts, has already been published. New versions will be released as more texts are added, approximately every six months (Stenroos and Mäkinen 2008: 2).

13 The inclusion of the normalisation of the text in the editing phase instead of the corpus compilation phase is based on the assumption that the editor of a historical text is usually more familiar with both the individual text and its linguistic conventions than the corpus linguist, who is dealing with a larger selection of potentially very different texts.

14 The MEG project also initially considered adopting TEI XML P5 as the annotation format, but due to reasons of convenience and compatibility opted at least initially for an encoding system based on that developed for LAEME (Stenroos and Mäkinen 2008: 6). According to Mäkinen (personal communication), moving over to XML at some later stage has not been ruled out and the encoding system used by the project has been kept such that it can be easily converted to XML at a later date.

15 The TEI Guidelines themselves are available under the terms and conditions of the GNU General Public License (version 2, <http://www.gnu.org/licenses/old-licenses/gpl-2.0.html>), which means that the DECL guidelines will also need to be published under a compatible free software license.

16 For an explanation of the terms of this license, see <http://creativecommons.org/licenses/by-nc-sa/3.0/>.

17 With the notable exception of the National Archives of the United Kingdom, which allows researchers to take digital photographs of their archival material.

18 Meurman-Solin's properties for fourth-generation corpora are much the same as the DECL principles of transparency, flexibility and expandability, combined with the cultivation of international standards and retaining the link to the manuscript reality. See also sections 2 and 5 above.

19 For more, and up-to-date, information, please visit the DECL website at http://www.helsinki.fi/varieng/domains/DECL.html.

Editions, corpora and related projects

Auchinleck Manuscript. <http://www.nls.uk/auchinleck>. Accessed 12 August 2008.

Boyle Papers Online. <http://www.bbk.ac.uk/boyle/boyle_papers/boylepapers_index.htm>. Accessed 12 August 2008.

British National Corpus (BNC). <http://www.natcorp.ox.ac.uk>. Accessed 18 August 2008.

Caxton's Canterbury Tales: The British Library Copies. (2003), B. Bordalejo (ed.) CD-ROM. Birmingham: Scholarly Digital Editions.

Centre for Scholarly Editing and Document Studies (Centrum voor Teksteditie en Bronnenstudie - CTB) at the Royal Academy of Dutch Language and Literature (KANTL) in Ghent, Belgium. <http://www.kantl.be/ctb>.

Corpus of Early English Correspondence (CEEC). (1998). Compiled by T. Nevalainen, H. Raumolin-Brunberg, J. Keränen, M. Nevala, A. Nurmi and M. Palander-Collin at the Department of English, University of Helsinki. Description available at <http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html>.

Corpus of Early English Correspondence Sampler (CEECS). (1998). Compiled by J. Keränen, M. Nevala, T. Nevalainen, A. Nurmi, M. Palander-Collin and H. Raumolin-Brunberg at the Department of English, University of Helsinki.

Corpus of Early Ontario English, 1776-1899 (CONTE). Being compiled by S. Dollinger at the University of British Columbia. See S. Dollinger (2006), 'Oh Canada! Towards the Corpus of Early Ontario English', in: A. Renouf and A. Kehoe (eds.) *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi. 7-26.

Corpus of Late Modern English Texts (CLMET). <http://perswww.kuleuven.be/~u0044428/clmet.htm>. Accessed 18 August 2008.

Corpus of Middle English Scientific Prose, A. (ACOMESP). Being compiled at the University of Málaga in collaboration with the University of Glasgow by A. Miranda García, D. Moreno Olalla, G.D. Caie, J. Calle Martin, M. Laura Esteban Sequra, N. Obegi Gallardo, S. González Fernández Corugedo and Teresa Marqués Aguado. <http://hunter.filosofia.uma.es/manuscripts>. Accessed 15 August 2008.

Corpus of Northern English Texts from Old to Early Modern English. Being compiled at the University of Seville. Described in G. Amores Carredano,

J. Fernández Cuesta and L. García-García (2008), 'Elaboration of an electronic corpus of northern English texts from Old to Early Modern English', paper presented at the *Sixth International Conference on Middle English*, 24-26 July 2008.

Corpus of Scottish Correpondence, 1500-1730 (CSC). Being compiled at the University of Helsinki by A. Meurman-Solin.

Creative Commons. <http://creativecommons.org/>.

Digital Archive of Letters in Flanders (DALF). <http://www.kantl.be/ctb/project/dalf>. Accessed 12 August 2008.

Digital Editions for Corpus Linguistics (DECL). <http://www.helsinki.fi/varieng/domains/DECL.html>.

Distributed Editions Initiative. Description available at <http://www.itsee.bham.ac.uk/DistributedEditions>. Accessed 12 August 2008.

Early English Books Online (EEBO). Available to subscribers at <http://eebo.chadwyck.com/home>. Accessed 13 August 2008.

Eighteenth Century Collections Online (ECCO). Available to subscribers at <http://galenet.galegroup.com/servlet/ECCO>. Accessed 13 August 2008.

English Witness Depositions 1560-1760: An Electronic Text Edition (EWD). Being compiled by M. Kytö, P. Grund and T. Walker. Description available at <http://www.engelska.uu.se/witness.pdf>. Accessed 18 August 2008.

Evellum. <http://www.evellum.com>. Accessed 12 August 2008.

Helsinki Corpus of English Texts (HC). (1991). Department of English, University of Helsinki. Compiled by M. Rissanen (project leader), M. Kytö (project secretary); L. Kahlas-Tarkka, M. Kilpiö (Old English); S. Nevanlinna, I. Taavitsainen (Middle English); T. Nevalainen, H. Raumolin-Brunberg (Early Modern English). Description available at <http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html>.

Hooke Folio Online. <http://webapps.qmul.ac.uk/cell/Hooke/Hooke.html>. Accessed 12 August 2008.

Institute for Textual Scholarship and Electronic Editing. <http://www.itsee.bham.ac.uk>.

Lampeter Corpus of Early Modern English Tracts. Manual available at <http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM>.

Letters of Clemency from the Chancery of Brittany. <http://nicole.dufournaud.net/remission>. Accessed 12 August 2008.

Linguistic Atlas of Early Middle English, 1150-1325 (LAEME). (2007). The University of Edinburgh. Compiled by M. Laing and R. Lass. <http://www.lel.ed.ac.uk/ihd/laeme/laeme.html>. Accessed 8 August 2008.

Linguistic Atlas of Older Scots, 1150-1325 (LAOS). The University of Edinburgh. Compiled by K. Williamson. <http://www.lel.ed.ac.uk/ihd/laos1/laos1_frames.html>. Accessed 18 August 2008.

London Provisioner's Chronicle, 1550-1563, by Henry Machyn: Manuscript, Transcription, and Modernization. Edited by R.W. Bailey, M. Miller and

C. Moore. <http://quod.lib.umich.edu/m/machyn>. Accessed 12 August 2008.

Middle English Medical Texts (MEMT). (2005). I. Taavitsainen, P. Pahta and M. Mäkinen (eds.) CD-ROM. Amsterdam: John Benjamins.

Middle English Grammar Corpus (MEG-C). (2008). Version 1.0. University of Stavanger. Compiled by M. Stenroos, M. Mäkinen, S. Horobin, J. Smith. http://www.uis.no/research/culture/the_middle_english_grammar_project. Accessed 6 August 2008.

Old Bailey Corpus (OBC). Coordinated by M. Huber. Description available at <http://www.uni-giessen.de/oldbaileycorpus/index.php>. Accessed 14 August 2008.

Papers of Sir Joseph Banks. <http://www2.sl.nsw.gov.au/banks/>.

Penn-Helsinki Parsed Corpus of Middle English (PPCME2). (2000). 2nd edition. A. Kroch and A. Taylor (eds.). Description available at <http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/>. Accessed 14 August 2008.

Proceedings of the Old Bailey, London's Central Criminal Court, 1674 to 1913. Old Bailey online. <http://www.oldbaileyonline.org>. Accessed 21 August 2008.

Representative Corpus of Historical English Registers, A (ARCHER). See D. Biber, E. Finegan and D. Atkinson (1994), 'ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers', in: U. Fries, G. Tottie and P. Schneider (eds.) *Creating and Using English Language Corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*, Zürich 1993. Amsterdam: Rodopi. 1-13. See also <http://www.anglistik.uni-freiburg.de/Englisches_Seminar/Lehrstuehle/LS_Mair/research/projects/archer/>.

Salem Witch Trials Documentary Archive and Transcription Project. <http://etext.virginia.edu/salem/witchcraft/>. Accessed 11 November 2008. For the Salem Witchcraft Records corpus, see P. Grund, M. Kytö and M. Rissanen (2004), 'Editing the Salem witchcraft records: an exploration of a linguistic treasury', *American Speech*, 79: 146-166.

Scholarly Digital Editions. <http://www.sd-editions.com>. Accessed 12 August 2008.

State Papers Online. <http://gale.cengage.co.uk/statepapers>. Accessed 5 February 2009.

State Papers Project. Coordinated by H. Good. <http://www.sp12.hull.ac.uk>. Accessed 12 August 2008.

Text Encoding Initiative (TEI). <http://www.tei-c.org>.

References

Bailey, R.W. (2004), 'The need for good texts: the case of Henry Machyn's Day Book, 1550-1563', in: A. Curzan and K. Emmons (eds.) *Studies in the History of the English Language II: Unfolding Conversations.* Topics in English Linguistics 45. Berlin and New York: Mouton de Gruyter. 217-228.

Curzan, A. and C.C. Palmer (2006), 'The importance of historical corpora, reliability, and reading', in: R. Facchinetti and M. Rissanen (eds.) *Corpus-based Studies of Diachronic English*. Bern: Peter Lang. 17-34.

De Smet, H. (2005), 'A corpus of Late Modern English texts', *ICAME Journal*, 29: 69-82. Available at <http://icame.uib.no/ij29>.

Deegan, M. and S. Tanner (2004), 'Conversion of primary sources', in: S. Schreibman, R. Siemens and J. Unsworth (eds.) *A Companion to Digital Humanities*. Malden: Blackwell Publishing. 488-504.

Dollinger, S. (2004), ''Philological computing' vs. 'philological outsourcing' and the compilation of historical corpora: a Late Modern English test case', *Vienna English Working Papers* (VIEWS), 13(2): 3-23.

Functional Requirements for Bibliographic Records (FRBR). Final report. Ifla study group on the functional requirements for bibliographic records (1998). Available at <http://www.ifla.org/VII/s13/frbr/frbr.htm>. Accessed 15 August 2008.

Grund, P. (2006), 'Manuscripts as sources for linguistic research: a methodological case study based on the Mirror of Lights', *Journal of English Linguistics*, 34: 105-125.

Huber, M. (2007), 'The Old Bailey proceedings, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English', in: A. Meurman-Solin and A. Nurmi (eds.) *Annotating Variation and Change.* Studies in Variation, Contacts and Change in English 1. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. <http://www.helsinki.fi/varieng/journal/volumes/01/huber>. Accessed 10 July 2008.

Kytö, M. (compiler) (1996), *Manual to the Diachronic Part of The Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*, 3rd edition. Helsinki: University of Helsinki Department of English. Available at <http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM>. Accessed 17 June 2008.

Kytö, M., P. Grund and T. Walker (2007), 'Regional variation and the language of English witness depositions 1560-1760: constructing a 'linguistic' edition in electronic form', in: P. Pahta, I. Taavitsainen, T. Nevalainen and J. Tyrkkö (eds.) *Towards Multimedia in Corpus Studies.* Studies in Variation, Contacts and Change in English 2. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. <http://www.helsinki.fi/varieng/journal/volumes/02/kyto_et_al>. Accessed 10 July 2008.

Lass, R. (2004), 'Ut custodiant litteras: editions, corpora and witnesshood', in: M. Dossena and R. Lass (eds.) *Methods and Data in English Historical Dialectology*. Linguistic Insights 16. Bern: Peter Lang. 21-48.

Machan, T.W. (1994), *Textual Criticism and Middle English Texts*. Charlottesville and London: University Press of Virginia.

Meurman-Solin, A. (2007), *Manual for the Corpus of Scottish Correspondence, 1500-1730*. <http://www.helsinki.fi/varieng/csc/manual>. Accessed 18 August 2008.

Nurmi, A. (1999), 'The Corpus of Early English Correspondence Sampler (CEECS)', *ICAME Journal*, 23: 53-64.

Rissanen, M. (2000), 'The world of English historical corpora: from Cædmon to computer age', *Journal of English Linguistics*, 28: 7-20.

Robinson, P. (2005), 'Current issues in making digital editions of medieval texts – or, do electronic scholarly editions have a future?', *Digital Medievalist*, 1.1 (Spring 2005). <http://www.digitalmedievalist.org>. Accessed 12 May 2008.

Shillingsburg, P. (1986), *Scholarly Editing in the Computer Age. Theory and Practice*. Athens, Georgia: The University of Georgia Press.

Stenroos, M. and M. Mäkinen (2008), 'Corpus manual, 1.0', in: *The Middle English Grammar Corpus*, M. Stenroos, M. Mäkinen, S. Horobin and J. Smith (compilers). Stavanger: University of Stavanger. <http://www.uis.no/getfile.php/Forskning/Kultur/MEG/Corpus manual 1.0.rtf>. Accessed 18 June 2008.

Vanhoutte, E. and R. Van den Branden (forthcoming), 'Describing, transcribing, encoding and editing modern correspondence material: a textbase approach', in: *Computing the Edition. Special Issue of Literary & Linguistic Computing*. Preprint available at <http://www.kantl.be/ctb/pub/preprint/comedvanvanfig.pdf>. Accessed 18 August 2008.

Parser-based analysis of syntax-lexis interactions

Hans Martin Lehmann and Gerold Schneider

University of Zurich

Abstract

In this paper we present a corpus-driven approach to the detection of syntax-lexis interactions. Our approach is based on the output of a syntactic parser. We have parsed the British National Corpus *and constructed a database of lexical dependencies. Such a large-scale approach allows for a detailed investigation of patterns and constructions associated with individual lexical items found in argument positions.*

We then address the methodological problems of such an approach: precision errors (unwanted instances) and recall errors (missed instances) and offer a detailed evaluation. We investigate the interaction between syntax and lexis in verb-subject and verb-object structures as well as the active-passive alternation. We show that our approach provides relatively clean data and allows for a corpus-driven investigation of rare collocations.

1. Introduction

Fixedness in language has been extensively studied in areas like multi-word units, idiomatic expressions, collocations and verb-particle constructions. These have often been treated as relatively fixed non-compositional sequences, which allow for little variation. In our paper, we will focus on co-occurrence phenomena between elements in syntactic relations. Specifically, we focus on subject-verb and verb-object relations in active and passive constructions. Looking for fixedness in these syntactic relations where compositionality is expected to hold to a large degree may strike the reader as a strange undertaking. Our main interest lies in establishing how far an open choice principle holds and to what degree we can find fixedness in these syntactic relations.

The identification of syntactic relations requires syntactically annotated corpora. Most standard corpora of sufficient size are either not annotated at all, or annotated at the non-hierarchical level of part-of-speech tags only. They typically contain no hierarchical information about the syntactic organisation of sentences. Due to the Zipfian nature of the frequency distribution of lexical items, smaller syntactically annotated corpora like the Penn Treebank (Marcus et al. 1993) and the British component of the *International Corpus of English* (ICE-GB, Nelson et al. 2002) impose serious restrictions on the study of syntax-lexis interactions.

Parsing approaches to fixedness are still quite rare. Exceptions are Lin (1998) and Seretan and Wehrli (2006). Robust broad-coverage syntactic parsers, for example Schneider (2008) or Andersen et al. (2008), have now become available, offering new perspectives on this type of research.

This paper describes the syntactic annotation of the 86.5 million running words in the written BNC with the help of Pro3Gres, a dependency parser.[1] We document the extraction of a database with verb centres and their dependents. We then explore the possibilities and limitations of this dependency database for the study of fixedness in syntactic relations.

2. Previous work

Most approaches to fixedness in language are based on the use of observation windows or regular expression patterns over large corpora with flat part-of-speech annotation. Typically, collocations and multi-word expressions are investigated. Intuitively, approaches based on syntactic hierarchical structure are to be expected to perform better on the task of recognising syntax-lexis interaction than surface-based approaches. Syntactic analysis has indeed been recognised as a prerequisite for accurately describing the syntax-lexis interface:

> Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure. However, in practice, free-style texts contain a great deal of nonstandard features over which automatic parsers would fail. This fact is being seriously challenged by current research [...] and might not be true in the near future. (Smadja 1993: 151)

The currently available corpora which are manually analysed for syntactic structure, for example ICE-GB and the Penn Treebank, are too small for infrequent word-word interactions, and automatic parsers have, until recently, not been robust enough to analyse large corpora. This partly explains why most approaches have been based on observation windows or part-of-speech sequences over corpora without hierarchical syntactic annotation.

Such surface-based approaches have made a wealth of corpus research possible, but some types of research can profit greatly from hierarchical syntactic information. Let us consider a simple example. Research on verb subcategorisation and selectional preferences needs to retrieve verb-object relations. The verb-object relation is one of the least problematic relations, since the distance between the verb and the object is quite short in English. In verb-subject relations, for example, relative clauses or appositions may intervene, which further complicates retrieval. In the following, we discuss which types of errors part-of-speech sequences and windows-based methods are typically prone to and justify our use of a parsing approach.

2.1 Part-of-speech tag sequences

Sequences of part-of-speech tags or regular expressions over part-of-speech tags (e.g. Hoffmann and Lehmann 2000; Heid and Weller 2008) have been used to describe collocations. Such search strategies may lead to various errors. Regular expression search strings will involve a verb tag followed by a noun tag, typically at some distance, and possibly limiting the context between the verb and the noun tag. Such a search will yield a wealth of examples. However, many of them will be incorrect (precision errors), and many verb-object relations will not be reported at all (recall errors).

- Precision errors: in sentences such as *Experts fear the Epstein Barr virus will spread* the regular expression will erroneously report a verb-object relation between *fear* and *virus*. In sentences like *The report arrived Friday* the regular expression will erroneously report a verb-object relation between *arrived* and *Friday*.
- Recall errors: in sentences such as *John likes swimming* the regular expression will not find the verb-object relation, because swimming is not a noun, but a verb (participle). In sentences like *John likes, but Mary hates Paul* the regular expression will probably not find the verb-object relation because the distance is relatively long, and the intervening part-of-speech tags are not the beginning of a noun phrase. Discarding such restrictions on intervening tags would probably lead to a precision error, erroneously reporting a verb-object relation between *likes* and *Mary*. In sentences such as *The potatoes I like are cold* a regular expression will not find the verb-object relation which is implicitly contained in the relative clause.

2.2 Windows-based approaches

Windows-based methods (e.g. Stubbs 1995) are still commonly used for collocation detection. N words before and after a key word (e.g. a verb) are considered. N is typically about three. The distinction between different types of collocations (e.g. subject-verb, verb-object and verb-PP) is typically left underspecified.

- Precision errors: in addition to the errors of the part-of-speech sequence method, windows-based methods suffer from precision errors due to the lack of implicit head extraction. In the example sentence *Experts fear the Epstein Barr virus will spread* the windows-based method also reports *fear Epstein* and *fear Barr* as collocation counts.
- Recall errors: recall is intrinsically low because many of the dependencies appear further than N words away. Recall can be increased by increasing N, but at the forbidding cost of decreasing precision.

2.3 Parsing approaches

Parsing approaches are still rarely used for investigating collocations, which may be partly related to the fact that some definitions of collocations, in contrast to

others, under-specify syntactic relations and are purely surface-based, for example as "sequences of lexical items that habitually co-occur" (Cruse 1986: 40). We take the view that syntax-lexis interactions are closely connected to individual syntactic functions and should abstract away from surface sequences as far as it is possible. We therefore base our investigation on syntactic functions. We present results on the subject and object function, but we hope to show that such an approach has a far wider potential.

A second major reason why parsing-based approaches are rare is that parsers which are sufficiently robust, fast and performant, for example Collins (1999), Schneider (2008) or Andersen et al. (2008), have only recently become available. Some of the few parsing approaches to collocation and multi-word expression (MWE) detection are Lin (1998) and Seretan and Wehrli (2006). Seretan and Wehrli (2006) have carefully evaluated their approach. They conclude that, in comparison to windows-based approaches, their parser-based system performs worse for the top-ranked collocations, but better in total.

> As for the MWE precision, the window method performs better for the first 100 pairs; on the remaining part, the parsing-based method is on average 3.7% better. The precision curve for the window method shows a more rapid degradation than it does for the other. Therefore we can conclude that parsing is especially advantageous if one investigates more than the first hundred results (as it seems reasonable for large extraction experiments). (Seretan and Wehrli 2006: 959)

Although we have not conducted an extensive evaluation of our approach as a collocation and MWE finder, our preliminary results support Seretan and Wehrli: some consistent tagging and parsing errors are ranked very high, but low count instances contain considerably less "garbage" than windows-based results typically show. The majority of nonce occurrences are syntactically correct. While parsers make errors, the number is low enough to profit collocation detection.

Two dependency types are especially hard to recover without parsing approaches: long-range dependencies and passive subjects. The majority of long-range dependencies span more than five words. Passive subjects are difficult for a number of reasons: (i) subject-verb distances are often much longer than verb-object dependencies; (ii) the recognition of passive forms by means of window methods or regular expressions is difficult; and (iii) passive verb forms are typically one word longer than active verb forms, introducing additional recall and verb-head extraction errors for windows-based methods. The passive subject is particularly interesting for syntax-lexis interaction: first, it undergoes selectional restrictions to the same degree as verb-object relations (it is an internal argument), and second, the passive is a marked construction (cf. Heid and Weller 2008).

2.4 Precision and recall of parsing approaches

Although parsing approaches perform better than surface-based approaches, the main challenge in using automatically parsed data remains, namely the fact that the data contain errors, which means that the analyses reported cannot simply be taken at face value. The success rate of automatic annotation tools is measured in terms of precision and recall. In an evaluation, one carefully annotates a small random selection (the so-called gold standard) and compares it to the automatic annotation. Precision reports how many of the automatically annotated instances are also contained in the gold standard, i.e. are correct. Recall reports how many of the instances in the gold standard are actually found in the automatic annotation.

Precision errors can be filtered relatively easily, unless the amount of data reported is too big, by going through the output of an automatic system and discarding the false positives, which are often referred to as unwanted instances or as garbage. Our data contains several million instances, so that filtering is not an option. Instead, we use the results of our evaluation as an estimate of the number of false positives that we need to expect from our tools.

Recall errors, also known as missed instances, are an even more serious problem. The only reliable way to know what the automatic annotation tool misses is to go through the entire corpus manually. The complete lack of information on what was missed makes it impossible to extrapolate from data based on automatic annotation. With a selective evaluation of the tool's performance – not just for overall performance but for the specific research question – the corpus linguist can, however, extrapolate from the missed instances in the gold standard to the new material processed in his/her research. We present a detailed evaluation of our tools' performance in section 4.

3. Data and method

In this section, we describe the data and the methods that we use. Our approach is based on a complete syntactic analysis of the *British National Corpus* (BNC). In our study, we use the written component of the BNC world-edition (cf. Aston and Burnard 1998) comprising approximately 86.5 million words.[2] Our processing pipeline consists of tagging, chunking, head-extracting and parsing the corpus. The parsed material is then imported and queried in a database. In the following, we describe the processing steps in detail, and illustrate each of them with the following example sentence:

(1) In the early days the stigma of being HIV positive had driven away about 60% of my circle of friends. (BNC A00: 189)

First, the corpus is tagged and the morphological base form, the lemma, is reported. For this step we used the decision-tree tagger Treetagger (Schmid

1994).[3] We chose to discard the part-of-speech tags included in some of the corpora, for example in the BNC, for reasons of consistency and cross-corpus comparability. Taggers assign morphosyntactic part-of-speech information to each word in the input text. We used the Penn Treebank Tagset (Marcus et al. 1993). After the first step, each word form is followed by the lemma and the tag, separated by underscores, the example sentence looks as follows:

(1a) In_in_IN the_the_DT early_early_JJ days_day_NNS the_the_DT stigma_stigma_NN of_of_IN being_be_VBG HIV_hiv_NNP positive_positive_NN had_have_VBD driven_drive_VBN away_away_RB about_about_IN 60%_CARD_CD of_of_IN my_my_PRP$ circle_circle_NN of_of_IN friends_friend_NNS

In the second step, noun groups and verb groups are recognised by means of a chunker. We use the conditional random fields chunker Carafe.[4] After the second step, verb groups and noun groups are marked by double square brackets, the example sentence looks as follows:

(1b) In_in_IN
[[the_the_DT early_early_JJ days_day_NNS]]
[[the_the_DT stigma_stigma_NN]]
of_of_IN being_be_VBG HIV_hiv_NNP positive_positive_NN
[[had_have_VBD driven_drive_VBN away_away_RB]]
about_about_IN
[[60%_CARD_CD of_of_IN my_my_PRP$ circle_circle_NN]] of_of_IN
friends_friend_NNS

In the third step, the heads of the chunks are extracted. The head of a verb chunk is typically the rightmost verb, the head of a noun chunk is typically the rightmost noun. Then, the corpus is syntactically analysed. The parser Pro3Gres, which we use, is dependency-based. It reports syntactic functions arranged in a tree structure. The parser is very fast and robust. It parses the entire BNC in little over 24 hours. It has been applied in many areas of research, for example information retrieval (Bayer et al. 2004), relation mining in biomedicine (Rinaldi et al. 2007) and psycholinguistics (Schneider et al. 2005). It was developed by one of the authors of this paper and is described in detail in Schneider (2008).[5] A screenshot of the graphical output of the dependency tree for the example sentence is given in figure 1 (with relations inside chunks excluded for simplicity).

The syntactic analysis of the example sentence conveys, for example, that *drive*, the head of the verb chunk *had driven away*, is the main verb which attaches a prepositional phrase (relation *pobj*) and has a subject (*subj*) and an object (*obj*). The object, which is headed by *circle*, is modified by a prepositional phrase (*modpp*). Inevitably, steps one to three of our method introduce a certain amount of error, which affects the results of our experiments. For a detailed analysis and evaluation see section 4.

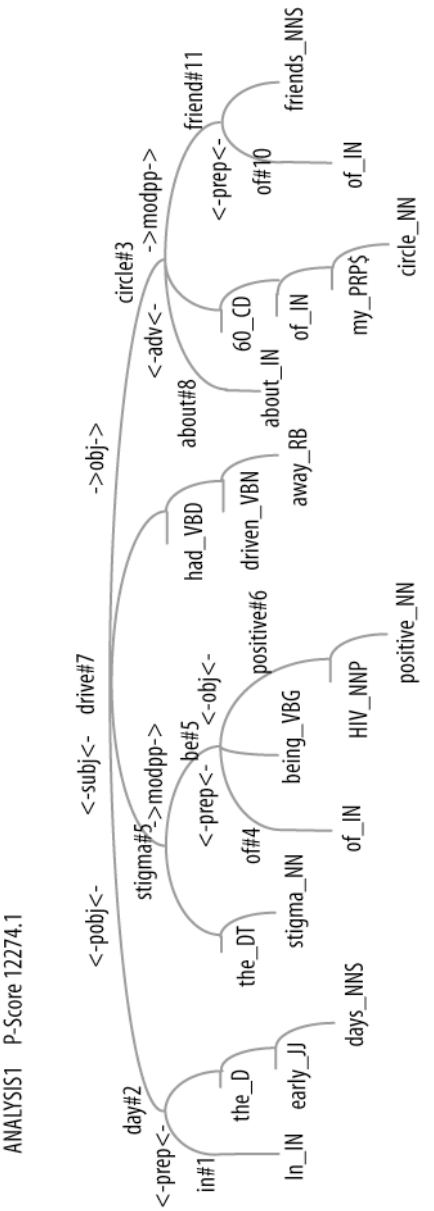

Figure 1: Output of the parser Pro3Gres for our sample sentence (BNC A00: 189)

The fourth step concerns accessing this richly annotated data. The parsed corpora were imported into a large database. We used Prolog to extract the selected data from the corpus and mySQL for storing the data. For each head of predicate identified by the parser, the database contains one record with cells describing the properties of the predicate head as well as its dependents, i.e. head of subject, head of object and PPs with preposition and description noun. For each of these we stored the word-form, the lemma, the part of speech, the direction of the dependency and the position in the sentence. In addition, the predicate was annotated for voice and finiteness. The analysis of the corpus results in a database with 10.5 million records for the written component of the BNC. This database forms the basis for our investigation described in section 5.

The choice of tools used for such a detailed linguistic investigation can have an impact on the results. Some of the presented ranked lists, for instance, are affected by tagging and parsing errors. Different taggers and parsers often make similar mistakes and have similar error rates, so that using a different tagger and parser will probably have little influence. Training a tagger over large manually annotated corpora from the individual domains would improve the results, but such corpora are not available yet.

While tagsets are standardised, different chunkers may follow different policies. The chunker that we use, Carafe, takes a very "greedy" and semantic approach, as we illustrate in the following. Chunkers typically return noun groups (which are typically unnested NPs) and verb groups. In the sentence

(2) The official spokesperson of Bogus Ltd. remained silent

we have the noun groups *the official spokesperson* and *Bogus Ltd* and the verb group *remained*. There are a number of syntactic configurations where different chunkers report different chunks, however. In the sentence

(3) One of the official spokespersons of Bogus Ltd. wanted to remain silent

our chunker reports the noun groups *one of the official spokespersons* and *Bogus Ltd* and the verb group *wanted to remain*, while many less greedy chunkers report *one* and *the official spokespersons* as separate noun groups and *wanted* and *remain* as separate verb groups. The greedy chunking option, which owes its name to the fact that the chunker greedily creates large chunks, typically coincides with being more semantic and less syntactic in nature. The greedy chunker reports a subject relation between *spokesperson* and *remain* in both sentences, abstracting away from surface syntax, while a non-greedy chunker would report a subject relation between *one* and *want* in sentence (3) (and probably a long-range dependency between *one* and *remain*). While the present investigation of lexical semantics supports a semantic chunking policy, research on modal verbs would warrant the use of a non-greedy chunker.

4. Parser evaluation and error handling

All automatic annotations face the problem that they are error-prone. Evaluation of the performance of automatic annotations is a major research topic in computational linguistics. We report in detail the error rates for the parser, including errors that stem from processing steps prior to parsing, such as tagging and chunking. First, we explain why an extensive evaluation of automatic tools is essential for corpus linguistics and quantitative language description in general. Second, we introduce a standard evaluation methodology. Third, we give a general evaluation of the parser and an evaluation of the relations that we have used in our research from sufficiently large random samples of the BNC. Finally, we discuss methods to cope with certain rates of error.

4.1 The need for extensive evaluation

The use of automatic taggers which introduce about 2-5 percent errors on average per token, depending on the tagset, the tagger, and the text type, is widespread in corpus linguistics, and this level of error rate is often tacitly acknowledged. While such a low error rate poses no problems to most frequency-based linguistic research, one needs to consider that errors may not be spread homogenously over the tagset. While some tags reach a performance of 99 percent, others may have a much lower performance. The distinction between prepositions (tag *IN* in the Penn tagset) and verbal particle (tag *RB* in the Penn tagset) is particularly difficult, because the context often looks identical. Some taggers achieve only 10 percent recall and 84 percent precision on this distinction (Baldwin and Villavicencio 2002). Research on verb particles which is based on tagged data may thus be seriously affected, despite the low error rate on average, per token.

A crucial step for assessing the effect that errors are thus causing is to carefully evaluate the performance of the used tools, not only on a general level, but particularly on the linguistic phenomena under investigation, and on the actual corpora. Evaluations that break down results by particular linguistic categories are also known as selective evaluations. Lin (1995) introduces the selective evaluation method for dependency parsers which we use here. An extensive selective evaluation allows one to extrapolate to the number of false positives (precision errors) and to the number of missed instances (recall errors) within reasonable limits.

4.2 Standard test corpus evaluations

A number of manually annotated corpora are standardly used to compare the performance of syntactic parsers. One of them is the GREVAL corpus (Carroll et al. 2003), which contains 500 near-random sentences from the *Suzanne Corpus*, covering a broad range of news texts. Performance of the Pro3Gres parser on the relations ‘subject’, ‘object’, ‘PP-attachment to verb’ and ‘PP-attachment to noun’ are given in table 1.[6]

Table 1: Performance of Pro3Gres on the 500 GREVAL sentences

Performance on GREVAL	Subject	Object	Noun-PP	Verb-PP
Precision	92%	89%	74%	72%
Recall	81%	84%	66%	84%

The parser has also been evaluated on one of its application areas, namely biomedical texts. 100 random sentences from the domain were manually annotated and compared to the parser output. The performance numbers are reported in table 2. An independent evaluation mapping Pro3Gres output to the Stanford dependency scheme was conducted by Haverinen et al. (2008), confirming the state-of-the-art performance of the parser.

Table 2: Performance of Pro3Gres on 100 random biomedical literature sentences

Performance on GENIA	Subject	Object	Noun-PP	Verb-PP
Precision	90%	93%	85%	82%
Recall	87%	91%	82%	84%

4.3 Evaluation on the BNC

The above evaluations show that for some of the argument structure relations, particularly subject and object, error rates lie between 10 and 20 percent. In the following, we give a selective evaluation of subject and object performance on BNC texts.

4.3.1 Subjects and objects

In order to test if these error rates carry over to our application corpora, for example the BNC, we manually annotated a small random selection of 100 spoken and 100 written sentences from the BNC. Performance results are given in table 3.

Table 3: Performance of Pro3Gres on 100 random sentences from the BNC

	BNC written		BNC spoken (con.gov.)	
	Percent	Count	Percent	Count
Subject precision	86%	108 / 125	88%	125 / 140
Subject recall	83%	108 / 130	89%	125 / 142
Object precision	87%	71 / 82	78%	70 / 90
Object recall	88%	71 / 80	87%	70 / 80

The results are similar to those obtained on the standard test corpora. Performance on the spoken corpus, particularly objects, is affected by our current rudimentary way of filtering hesitation markers (*errrm* etc.) and can be expected to improve with a better filtering algorithm.

4.3.2 Passive subjects

One of the applications that we will discuss in section 5 involves passive subjects, a subgroup of subjects that has special characteristics and may show a different performance. A separate evaluation is thus appropriate. The 100 random sentences from the BNC contained only five passive subjects in the spoken and 14 in the written part, these counts are too low to allow a reliable evaluation. In order to attain sufficiently large counts, 100 random sentences from the written BNC which contain verb participles (Penn tag *VBN*) were manually annotated for passive subjects and passive verb forms. The performance thus found is given in table 4. The passive subject error rate (left column) is similar to the general subject error rate, although slightly lower. Some passive subject errors are due to the fact that our passive verb form recognition algorithm has only 91 percent precision and 92 percent recall (right column). Since our random selection method of using only sentences containing a *VBN* tag tacitly assumes that the tag *VBN* is always correct, passive subject performance can, in reality, be expected to be 1-4 percent below the figures in table 4.

Table 4: Passive subject evaluation, based on 100 random BNC sentences containing verb participles

	Passive subject BNC written		Passive verb forms BNC written	
	Percent	Count	Percent	Count
Precision	85%	58 / 68	91%	60 / 66
Recall	82%	58 / 71	92%	60 / 65

4.3.3 Local and nonlocal subjects and objects

We have hitherto assumed that everybody knows what subjects and objects are. Although we use standard terminology, a definition is called for. We use the term 'subject' to either denote the explicit subject of a finite verb, or the implicit subject of an infinite or finite verb. Implicit subjects of finite verbs are relative pronoun resolutions (for example, *Girls who like boys*, where *girls* is the implicit subject of *like*), and implicit subjects of infinite verbs are control structures (for example, *Peter is unable to win*, where *Peter* is the implicit subject of *win*). Other types of implicit subjects, such as pronoun resolution (for example, *Peter sleeps and he snores*, where *he* is implicitly *Peter*), or indexed gerunds (for example, *Peter entered, cheering*, where the implicit subject of *cheer* is *Peter*) are not returned by our parser. The definition of objects is analogous.

The implicit subjects and objects reported by our parser are so-called non-local dependencies. Non-local dependencies are also termed long-distance or long-range dependencies. For example, in sentence (4), *procedure* is the implicit subject of the verb *transfer*, and *value* is the implicit subject of *indicate*.

(4) The procedure does not wait for offline modules to be transferred, however a value is returned in MODULES-ONLINE to indicate whether any modules are offline awaiting transfer. (BNC HWF: 4093)

In this case, the non-local dependencies are so-called subject-control relations. Non-local dependencies are more difficult to detect by automatic parsers. Separate performance values on the 100 BNC written random sentences (see table 3) broken down into non-local and local relations are given in table 5. While the counts on non-local relations are too low to deliver reliable results, local subject and object relations are parsed with almost 90 percent precision and recall.

Table 5: Performance of local and non-local relations on 100 sentences from the BNC

	Local relations BNC written		Non-local relations BNC written	
	Percent	Count	Percent	Count
Subject precision	89%	102 / 104	55%	6 / 11
Subject recall	86%	102 / 119	55%	6 / 11
Object precision	89%	70 / 79	33%	1 / 3
Object recall	89%	70 / 79	100%	1 / 1

4.4 Error handling

We have consistently found error levels between 10 and 20 percent for the subject and object relations, both in the standard evaluation corpora as well as in actual BNC data. While such error levels are too high to e.g. report absolute numbers reliably, we suggest that, based on a careful selective evaluation, limited scientific statements are possible, and that it is possible to quantitatively extrapolate to false positives (garbage, precision errors) and to missed instances (recall errors) within reasonable limits.

5. Exploring the syntax-lexis interface

In section 2 we described the extraction of a database containing verbs and their dependent subjects, objects and PPs. In this section we explore the wealth of data contained in these databases. We will focus on verb predicates and their subject and object dependencies.[7] The main interest driving our research is a quantitative analysis of the interaction between syntactic structures and the lexicon. We

extract and measure lexical preferences in the cline from free choice to collocation and structural preferences in the active-passive alternation using customised databases and statistical measures of surprise.

Any study of the interaction between lexical choices and syntactic choices in subject and object NPs and their governing verbs will have to take into account the active-passive alternation. Lexical choices will heavily depend on thematic roles. The use of parsed data allows us to deal with subjects of active verbal constructions separately from subjects in passive constructions. In the same way, we can limit our observations to objects in active constructions.

Windows-based approaches to syntax-lexis interaction (e.g. Stubbs 1995) take into account all content words present in the vicinity of each other (inside the observation window), irrespective of their syntactic function, and irrespective of whether they are syntactically connected at all. As a consequence, such non-hierarchical approaches are forced to base the expected value (E, null hypothesis) on the assumption of a corpus in which the words appear in random order. As Evert (2009) points out, such a null hypothesis is not unproblematic:

> [T]he null hypothesis of independence is extremely unrealistic. Words are never combined at random in natural language, being subject to a variety of syntactic, semantic and lexical restrictions. For a large corpus, even a small deviation from the null hypothesis may lead to highly significant rejection and inflated association scores calculated by significance measures. Effect-size measures are also subject to this problem and will produce inflated scores, e.g. for two rare words that always occur near each other (such as *déjà* and *vu*). A possible solution would be to specify a more realistic null hypothesis that takes some of the restrictions on word combinatorics into account, but research along these lines is still at a very early stage. (Evert 2009)

For these reasons, we avoid a null hypothesis (E) based on a random shuffling of all words in the corpus. Instead, we base our expectation on a random shuffling of the head verbs, and head nouns actually observed in the subject-verb and verb-object dependencies in our data. This offers a more realistic expectation by taking the syntactic restrictions on word combinatorics into account. The observed lexical head-verb preferences inside a fixed structure allow us to investigate selectional preferences and native-like selection (Pawley and Syder 1983). Comparing the use of the same head verbs and head nouns across different syntactic variants also permits us to investigate alternation preferences.

Our approach has the advantage of focusing on the construction under examination, for example verb-object relations ignoring frequent lexical items found in PPs and other constructions, which would skew our results otherwise. We use Observed divided by Expected (O/E) in the following because it has a clear probabilistic definition and is directly related to information-theoretic measures of surprise such as mutual information.

Unlike many other approaches, we use the lemma and not the word-forms in producing generalisations about the distribution of lexical items. On the one hand, this introduces a higher degree of generalisation; on the other hand, we may miss some interesting phenomena concerned with the co-occurrence of word-forms. The present empirical setup would easily allow for an investigation based on word-forms. However, in this paper we focus on the lemmas for establishing types of structural co-occurrence.

In order to avoid subject complements, all our calculations exclude the lemmas *be* and *seem* as head of predicate. Obviously, this is a crude approximation and should be incorporated in a more complete and systematic form in a future verb-valency database.

In the case of subject-verb combinations, we calculate the measure of surprise on the basis of a random shuffling of all subject heads and predicate heads found in the analysed corpus. This will result in an expected frequency for individual subject-verb combinations, E. The observed frequency of each subject-verb combination, O, is then used to calculate the measure of surprise, O/E, which expresses the factor by which the actual occurrence of the combination exceeds the expected frequency. Given the focus on subject-verb combinations, the measure of surprise here does not express the surprise of finding individual lexemes in subject position in general. The expectation is based on the assumption of free combination within the limits of the syntactic construction, i.e. any subject head can combine with any predicate head. In other words, E expresses free selection as opposed to selectional restriction.

Table 6 below shows the tendency of subject heads to co-occur with certain verb heads in subject-verb constructions. While, as linguists, we will not be surprised by the fact that *dogs bark*, *phones ring* and *thieves steal*, this data reminds us that selectional restrictions not only occur with internal arguments, but also with external arguments. *The phone rings* may be a purely semantic restriction, there are few synonyms to *ring*, and using them leads to a different meaning (e.g. *the phone tolls*, *the phone peals*). But *relations deteriorate* is largely a syntactic, idiomatic restriction, since *relations worsen*, *relations degenerate* have a similar meaning but remain rarer. Or, the verbs in *the sun shines*, *the sun beams* or *the sun radiates* have similar meanings but occur only rarely.

As we have shown in section 4, the annotation process is far from infallible. The combination *tentacle pore* is a case where the tagger failed and assigned a verb reading to *pores* in the compound *tentacle pores*. The combination *onion chop* is produced by a parsing error that failed to interpret postmodification by participle in sentences like (5), found in a recipe.

(5) 1 medium onion, chopped. (BNC BPG: 1001)

Sentences (6) and (7) show that the parser not only introduces errors but also contributes considerably to the coverage of our approach by including long-distance dependencies.

(6) For a number of reasons, however, the *eggs* failed to *hatch*. (BNC AM2: 103)
(7) Here the female may produce 800,000 *eggs* which *hatch* within 36 hours into larvae. (BNC A3Y: 43)

Table 6: Selectional restriction in active subject-verb combinations, f(sub_verb) > 50; BNC-world written component

subject_verb heads	f(sub_verb)	f(subject)	f(verb)	O/E
tentacle_pore	77	136	243	87.8576e+09
onion_chop	56	139	776	19.577e+09
egg_hatch	58	396	456	12.1116e+09
doorbell_ring	65	80	4220	7.26013e+09
interview(s)_record	136	136	5389	6.99722e+09
bomb_explode	158	652	1326	6.89128e+09
rumour_circulate	56	402	848	6.19441e+09
telephone_ring	195	356	4220	4.89447e+09
dog_bark	81	1739	506	3.47111e+09
phone_ring	142	374	4220	3.39264e+09
relation_deteriorate	62	950	738	3.33461e+09
sun_shine	294	1981	1690	3.31139e+09
lip_part	63	825	872	3.3022e+09
lifespan_display	111	420	3391	2.93886e+09
god_bless	135	3349	603	2.52078e+09
wind_blow	239	1381	2986	2.18549e+09
thief_steal	103	590	3015	2.18339e+09

The list in table 6 above has to be seen as raw material for closer analysis. It could, for instance, be used for extracting dictionary entries together with typical examples.

Table 7 below shows the association between subject and verb in passive constructions; *seeds are sowed*, *shots are fired* and *lessons learnt*. In terms of the active-passive alternation it is more interesting to compare passive subjects with active objects than with active subjects.

Table 7: Selectional restriction in passive subject-verb combinations, f(sub_verb) > 50; BNC-world written component

subject_verb heads	f(sub_verb)	f(subject)	f(verb)	O/E
seed_sow	51	172	147	16.762e+08
shot_fire	89	233	474	6.69664e+08
lesson_learn	73	217	604	4.62836e+08
duty_owe	59	431	282	4.03391e+08
battle_fight	55	235	521	3.733e+08
breakfast_serve	55	113	1375	2.94159e+08
offence_commit	193	398	1459	2.76198e+08
warrant_issue	73	150	1511	2.67651e+08
day_number	62	989	207	2.51667e+08
power_vest	124	1564	274	2.40456e+08
battle_win	53	235	788	2.37839e+08
prise_award	72	192	1345	2.31691e+08
attention_focus	149	1145	486	2.22508e+08
war_fight	57	427	521	2.12917e+08
treaty_sign	92	330	1160	1.99718e+08
reliance_place	52	69	3175	1.97248e+08
car_park	68	1121	263	1.91668e+08
interview_conduct	60	232	1189	1.80752e+08

Table 8 below shows the top 18 verb-object combinations. The combinations *inter alia*, *rick sky* and *programme tdy* are reported due to tagging errors. Rather than as a result in itself, this list can serve as raw material and a starting point for further investigation. For tasks such as lexicon construction, it is easy to filter lists of types containing a few systematic errors. We also find similar items to the ones found in table 7; e.g. *sow seeds* vs. *seeds are sowed*, where we find both alternations, active and passive.

Due to the size of the corpus and the extended coverage provided by the parser it is also possible to investigate the semantic prosody of low frequency items. Taking the combinations *wreak havoc* and *extol virtue* as a starting point, table 9 below shows the semantic prosodies of *extol* and *wreak*.

The dependency database developed for the purpose of this study thus allows for the exploration of phenomena like semantic prosody. It also provides links to the original corpus. Sentence (8) shows an example where the semantic prosody of *wreak* is used for creating a contrary effect.

(8) David Baddiel extolling hardcore porn? (TLN955826475)

Table 9, as with all our co-occurrence tables, is truncated, but otherwise unedited. Unlike the other tables it contains nonce occurrences. Our approach, based on parsed corpora, provides astonishingly clean data even at extremely low levels of

frequency. Given such promising results, we see the potential for investigating phenomena occurring at extremely low frequencies.

Table 8: Selectional restriction in active verb-object combinations, f(sub_verb) > 50; BNC-world written component

verb-object heads	f(sverb_obj)	f(verb)	f(subject)	O/E
inter_alia	259	348	269	8.26219e+10
wreak_havoc	88	157	248	6.7493e+10
whet_appetite	70	85	419	5.86938e+10
rick_sky	83	91	500	5.44746e+10
extol_virtue	56	132	379	3.34274e+10
programme_tdy	55	1068	55	2.79612e+10
clench_fist	82	399	403	1.52287e+10
beg_pardon	145	1320	216	1.51868e+10
grit_tooth	146	227	1363	1.40915e+10
purse_lip	135	184	1680	1.30417e+10
wrinkle_nose	82	202	1039	1.16674e+10
bridge_gap	162	321	1367	1.10247e+10
sow_seed	107	469	637	1.06954e+10
heave_sigh	74	512	430	1.00374e+10
buck_trend	58	184	1004	9.3757e+09
enclose_sae	68	1148	211	8.38324e+09
ratify_treaty	99	419	859	8.21401e+09
reap_reward	73	394	680	8.13664e+09

Table 9: Semantic prosody of *extol* and *wreak*

extol		wreak	
verb-object	n	verb-object	n
extol_virtue	56	wreak_havoc	88
extol_benefit	5	wreak_vengeance	10
extol_beauty	4	wreak_revenge	9
extol_man	2	wreak_destruction	5
extol_courage	1	wreak_damage	4
extol_approach	1	wreak_kind	2
extol_brilliance	1	wreak_mayhem	2
extol_authority	1	wreak_assault	1
extol_success	1	wreak_distortion	1
extol_achievement	1	wreak_pain	1
extol_leader	1	wreak_carnage	1
extol_riches	1	wreak_spite	1

It is typically recommended that O/E, the statistical measure that we have used here, be avoided because it has the tendency to rank as very high combinations where both words are rare. In windows-based and tag-sequence based approaches, where precision is typically problematic, this has the undesirable side effect that false positives dominate a large area at the top of the lists. Our parser-based approach suffers from this to a much lesser degree, opening up new possibilities for investigating combinations of rare words.

Table 10 shows the measure of surprise of finding subject-verb-object triplets. It was calculated from the data presented in tables 7 and 8. We filtered out occurrences of *page omitted advertisement* and *page omitted photograph*, which quite obviously are due to coding errors in the BNC, where they are not consistently set off as encoding comments.

Table 10: Fixedness in active subject-verb-object combinations, f(svo) > 20; BNC-world written component

subject-verb-object heads	f(svo)	f(s)	f(v)	f(o)	O/E
coroner_record_verdict	33	284	5389	517	60.8756e+11
spine_form_fan	23	113	12659	547	42.9051e+11
heart_miss_beat	26	1590	7393	222	14.5428e+11
clause_exclude_liability	25	744	3061	1126	14.2302e+11
jury_return_verdict	45	669	16513	517	11.5005e+11
sale_start_monday	23	1667	17732	159	7.14306e+11
female_lay_egg	29	683	6985	1317	6.73708e+11
sale_start_december	31	1667	17732	250	6.12315e+11
republic_achieve_independence	24	596	11145	1130	4.66717e+11
error_occur_error	21	583	13564	961	4.03354e+11
court_grant_injunction	22	7229	3247	345	3.96543e+11
inc_report_profit	145	1802	11740	2711	3.69031e+11
plc_report_profit	22	288	11740	2711	3.50332e+11
index_close_point	204	947	13999	8451	2.65780e+11
tenant_pay_rent	21	865	23676	666	2.24733e+11
corp_report_profit	65	1380	11740	2711	2.16015e+11
price_include_breakfast	195	3462	45097	948	1.92308e+11
history_repeat_itself	29	1186	3841	6172	1.50553e+11

Such triplets would be extremely difficult to retrieve with window-based or pattern based approaches and the low number of instances found in a 86.5 million-word corpus shows the advantage of our parser-based retrieval. Instances like sentence (9), which was retrieved by our methodology, are extremely difficult to locate with non-hierarchical strategies.

(9) It was a foregone conclusion that the *jury*, carefully selected beforehand, would *return* their immediate and unanimous *verdict* of “Guilty”. (BNC ALK: 796)

In the following, we focus on the study of the active-passive alternation and its interaction with lexical choices. We decided to investigate cases where the same pair of lexical items is involved in active as well as passive constructions as in *sow seeds* vs. *seeds are sowed*. Table 11 shows a ranking of such pairs ordered according to their preference for passive constructions.

Table 11: Preference for passive constructions for word pairs occurring in alternation in the written BNC; f(active)>2, f(passive>2, f(total)> 100

pair of lemmas	f(active)	f(passive)	f(total)	% passive
baby_bear	3	141	144	97.9167
study_carry	4	118	122	96.7213
committee_set	5	137	142	96.4789
power_vest	6	124	130	95.3846
test_carry	7	100	107	93.4579
research_carry	10	125	135	92.5926
system_base	10	106	116	91.3793
work_carry	29	274	303	90.4290
example_show	47	253	300	84.3333
case_adjourn	23	94	117	80.3419
election_hold	112	442	554	79.7834
people_arrest	31	113	144	78.4722
detail_obtain	33	105	138	76.0870
decision_base	45	118	163	72.3926
people_injure	45	102	147	69.3878
detail_find	43	95	138	68.8406
service_hold	38	78	116	67.2414
soldier_kill	38	75	113	66.3717

Table 11 shows the top 18 word pairs in terms of preference for the passive.[8] The restriction to pairs occurring more than 100 times ensures a minimal number of observations for the comparison between active and passive. The restriction to pairs that occur at least three times in the active as well as in the passive is applied in order to limit our observation to pairs for which the active-passive alternation is relevant.

The top-ranked pair *baby* and *bear* shows a massive preference for the passive. More than 97 percent of all observations occur in the passive, as in sentence (11). Active constructions, as in sentence (10), are extremely rare.

(10) They say drug-abusing mothers who would previously have had an abortion are *bearing* sickly *babies* with low chances of survival. (BNC A1G: 487)

(11) Overall, three in ten *babies* are *born* outside marriage in the UK to mothers of all age groups. (BNC K3S: 91)

The reported instances at the top of the list show a marked difference for the distribution of active and passive constructions, which is found at a level of 6-13 percent in our data. However, we have to take into account the restriction of our observation to 100 occurrences in total, with at least three occurrences for both variants. Combinations that occur only in the active or only in the passive voice are excluded. Based on the average of the pairs actually observed in table 11 we expect 16 percent of the instances realised as passives. Given the range of observed passive percentages from 98 percent for *bear baby* down to 0.6 percent for *make sense*, we can observe a strong interaction between active-passive constructions and lexical choices. The fact that in table 11 we only consider the middle ground between pairs that exclusively occur in either the active or the passive voice makes this observation more remarkable. To complete the picture at both ends of the cline, we present the pairs at the extreme ends in table 12.

Table 12: Active verb-object pairs that do not have a passive counterpart and passive verb-subject pairs that do not have an active counterpart

exclusively active		exclusively passive	
object verb	n	verb subject	n
time_have	3620	scroll_area	45
chance_have	2106	situate_hotel	43
power_have	1980	approve_study	43
way_go	1456	know_little	39
difficulty_have	1434	bear_william	38
interest_have	1353	base_figure	36
access_have	1229	base_diagnosis	34
opportunity_have	1228	base_some	32
lot_have	1207	call_fireman	30
impact_have	1185	enter_correspondence	29
reason_have	1100	age_cent	28
home_come	1065	set_council	28
choice_have	1017	wind_company	28
role_have	972	hand_judgment	25
money_have	970	situate_house	24
look_have	949	bear_george	24
sense_have	944	bear_thomas	23
implication_have	889	announce_date	23
influence_have	881	bear_james	22
job_get	851	suspend_share	22

Not surprisingly, we find the middle verb *have* dominating the most frequent combinations not occurring in the passive voice. The combinations exclusively occurring in the passive are more varied and occur at very low frequencies. Besides *be based on* we find *be born*, *be situated*, *be suspended* and *be an-*

nounced. Of course, such passive verbs can be found in the active voice in our data, as shown in sentence (12) for *be situated*.

(12) If you *situate* the *cable* tidy as close to your tank as possible, you can cut the wires on your equipment fairly short, to get rid of all those unsightly trailing cables. (BNC C97: 1725)

A quick glance through the type list of objects occurring with *situate* showed no hits for an object of the type 'building', whereas a type list for the passive subject immediately reveals *hotel*, *house*, *village*, *premise*, *property*, *school*, *office*, *station*, *centre*, *college* etc.

In table 13 we list the verbs occurring in the combinations presented in table 12 in abstraction from the objects or subjects they occurred with.

Table 13: Frequency of verbs in verb-object combinations not occurring in the passive and of verbs in subject-verb combinations not occurring in the active voice in the written component of the BNC

exclusively active		exclusively passive	
object verb	n	verb subject	n
have	136302	base	353
get	25436	bear	260
become	14054	associate	154
see	8305	situate	144
give	6172	set	123
do	5787	know	73
include	5090	approve	63
take	4983	deal	53
make	4672	scroll	45
want	4452	confine	44
come	3847	carry	41
feel	3813	concern	39
receive	3308	account	35
go	3053	hold	32
like	2811	call	30
meet	2708	enter	29
follow	2568	injure	29
show	2542	aim	28
leave	2521	age	28
allow	2438	wind	28

Among the verbs occurring in exclusively active combinations we find verbs that do not form the passive in general. These are characterised by a bundle of features like stative vs. dynamic, agent subject vs. non-agent subject etc. However, we also find verbs with asymmetric preferences for passive subjects

and active objects. Such asymmetries are due to a variety of causes. A closer analysis of these could be used for creating a corpus-driven classification of verbs. Given the highly problematic nature of terminology like 'middle verb' and 'light verb', such an empirical approach appears promising.

In some cases, the view tacitly taken here that we observe the active-passive alternations as a system of only two possibilities is wrong. Often we find semantically close variants, e.g. *decision be made* vs. *make decision* vs. *decide*. The verbs occurring in exclusively active combination contain many semantically weak verbs. We can see that semantically weak verb combinations are subject to restricted flexibility.

6. Discussion

All the results presented here are of an exploratory nature. Any of the phenomena explored could and should be studied in a more detailed analysis. As shown, the parser-based approach may help us in the study of identified co-occurrences, like *charge make*. However, the main impact of studying fixedness in language and the interface between syntax and the lexicon consists in the new possibility of a mainly corpus-driven rationale for the selection of individual co-occurrences in syntactic structures. While, as corpus linguists, we can easily explore identified co-occurrences, we were largely limited to approaches with a lexical node for studying these. The selection of a lexical node itself was motivated by hunches, intuition and previous work done in the field of study.

Pattern-based approaches may reach a similar or better recall rate in the identification of specific phenomena (see e.g. Lehmann 1997). Without massive manual intervention, however, they tend to incur an inacceptable level of precision, which severely limits the usefulness of the results. The only alternative for producing a more reliable ranking of the combinations is manual annotation, which for the written BNC would imply the manual creation and annotation of a database of more than ten million records.

The selection of large corpora is not only a pretext for using a parser. The precariously low numbers in the cells of our table 9 clearly show the necessity of analysing large corpora for this type of study. Our approach, based on automatically parsed corpora, provides astonishingly clean data even at extremely low levels of frequency. These promising results allow linguists to investigate phenomena occurring at extremely low frequencies.

Typically it is recommended that using O/E, the probabilistic measure that we have used here, be avoided because it has the tendency to rank as very high combinations where both words are rare. In windows-based and tag-sequence based approaches, this has the undesirable side effect that false positives dominate a large area at the top of the lists. The parser-based approach suffers from this to a much lesser degree, opening up new possibilities for investigating combinations of rare words.

The problem of recall has only been partly addressed by our present approach. The assumption that recall problems are distributed evenly across lexical types does not necessarily hold. For the future, we see great potential in making use of smaller manually annotated corpora like ICE-GB for testing such assumptions. Many of our results were created with an arbitrary cut-off at 50 or 100 instances. We are investigating statistical means for justifying such restrictions. From a statistical point of view this would pose no special problems. However, we have found that there is a lack of agreed degrees of certainty for such an undertaking. There is clearly a need for discussion in the field which would result in generally accepted significance levels.

7. Conclusions and outlook

In this paper we have described the compilation of a verb dependency database and shown its potential in several areas of research. We have presented a method for using partly erroneous parser output. This method is based on a selective evaluation which serves as a basis for extrapolating precision and recall errors.

We have outlined several applications exploiting the data compiled for this study. We have explored selectional preferences in subject-verb and verb-object combinations and investigated the active-passive alternation. We have discussed verb-subject collocations, although they have been investigated less vigorously than verb-object collocations. For verb-object relations, we have shown that our approach is well suited to the exploration of semantic prosody, and also of rare words.

For the active-passive alternation we have described the gradient area where lexical choices coincide with preferences for the active and passive construction, respectively. In the case of the active-passive alternation, our approach allowed us to quantify the gradient lexico-grammatical phenomena at the interface between syntax and lexis on a largely corpus-driven basis. Our investigation offers a corpus-driven rationale for locating syntax-lexis interactions instead of relying on intuition-based testing of collocations for individual words.

Currently, we are working on a more detailed database of verb-valency. We are including additional syntactic relations and their properties in order to offer a more detailed description of English argument structure. We are also exploring the possibility of comparing these preferences in different varieties of English.

Besides the exploitation of the database for lexicographical purposes, like rich lexicon entries indicating combinatorial preferences, we see the potential of the database in corpus-driven studies of other alternations, and the extraction and empirical description of special, theoretically problematic classes like middle verbs or ergative verbs as future applications.

Notes

1 See Schneider (2008) for a more detailed description.

2 Word counts for all corpora are based on a token-count of the tagged corpora excluding punctuation. These numbers may differ from those achieved with other approaches. However, due to the consistency of the counting method across all our data, they form a solid basis for comparison.

3 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

4 http://sourceforge.net/projects/carafe

5 http://www.ifi.uzh.ch/cl/gschneid/parser/

6 For more detailed evaluations, including a complete mapping to the relation set used in GREVAL and a comparison with several other syntactic parsers, refer to Schneider (2008).

7 We are aware of the intriguing possibilities offered by the dependent PPs stored with preposition and description noun in our database. However, we feel that such an analysis is beyond the scope of this paper.

8 The results in table 11 also reflect our conservative approach to multi word entities. Phrasal verbs and their particles are analysed as separate word tokens. As a consequence, the predicate head *carry* represents both *carry out* as well as *carry forward*.

References

Andersen, Ø.E., J. Nioche, T. Briscoe and J. Carroll (2008), 'The BNC parsed with RASP4UIMA', in: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Aston, G. and L. Burnard (1998), *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baldwin, T. and A. Villavicencio (2002), 'Extracting the unextractable: a case study on verb-particles', in: *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*. Taipei, Taiwan. 98-104.

Bayer, S., J. Burger, W. Greiff and B. Wellner (2004), 'The MITRE logical form generation system', in: *Proceedings of Senseval-3: The Third International Workshop on Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain.

Carroll, J., G. Minnen and E. Briscoe (2003), 'Parser evaluation: using a grammatical relation annotation scheme', in: A. Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer. 299-316.

Collins, M. (1999), *Head-driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, Philadelphia.

Cruse, D.A. (1986), *Lexical Semantics*. Cambridge: Cambridge University Press.

Evert, S. (2009), 'Corpora and collocations', in: A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics*: *An International Handbook*, article 58. Berlin: Mouton de Gruyter.

Haverinen, K., F. Ginter, S. Pyysalo and T. Salakoski (2008), 'Accurate conversion of dependency parses: targeting the Stanford scheme', in: *Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, Turku, Finland.

Heid, U. and M. Weller (2008), 'Tools for collocation extraction: preferences for active vs. passive', in: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Hoffmann, S. and H.M. Lehmann (2000), 'Collocational evidence from the *British National Corpus*', in: J. Kirk (ed.) *Corpora Galore. Analyses and Techniques in Describing English*. Amsterdam and Atlanta: Rodopi. 17-32.

Lehmann, H.M. (1997), 'Automatic retrieval of zero elements in a computerised corpus', in: M. Ljung (ed.) *Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi. 179-194.

Lin, D. (1995), 'A dependency-based method for evaluating broad-coverage parsers', in: *Proceedings of IJCAI-95*, Montreal.

Lin, D. (1998), 'Extracting collocations from text corpora', in: *Proceedings of First Workshop on Computational Terminology*, Montreal, Canada, 57-63.

Marcus, M., B. Santorini, and M. Marcinkiewicz. (1993), 'Building a large annotated corpus of English: the Penn Treebank', *Computational Linguistics*, 19(2): 313-330.

Nelson, G., S. Wallis and B. Aarts (2002), *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Pawley, A. and F. H.- Syder, (1983), 'Two puzzles for linguistic theory: native-like selection and native-like fluency', in: J.C. Richards and R.W. Schmidt (eds.) *Language and Communication.* London: Longman. 191-226.

Rinaldi, F., G. Schneider, K. Kaljurand, M. Hess, C. Andronis, O. Konstanti and A. Persidis (2007), 'Mining of functional relations between genes and proteins over biomedical scientific literature using a deep-linguistic approach', in: *Journal of Artificial Intelligence in Medicine*, 39: 127-136.

Schmid, H. (1994), 'Probabilistic part-of-speech tagging using decision trees', in: *Proceedings of International Conference on New Methods in Language Processing*.

Schneider, G. (2008), *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral thesis, Institute of Computational Linguistics, University of Zurich.

Schneider, G., F. Rinaldi, K. Kaljurand and M. Hess (2005), 'Closing the gap: cognitively adequate, fast broad-coverage grammatical role parsing', in: *Proceedings of ICEIS Workshop on Natural Language Understanding and Cognitive Science (NLUCS 2005)*, Miami, Florida.

Seretan, V. and E. Wehrli (2006), 'Accurate collocation extraction using a multilingual parser', in: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July. Association for Computational Linguistics. 953-960.

Smadja, F. (1993), 'Retrieving collocations from text: Xtract', *Computational Linguistics*, 19(1): 143-177.

Stubbs, M. (1995), 'Collocations and semantic profiles: on the cause of the trouble with quantitative studies', *Functions of Language*, 2(1): 23-55.

Index

D

Robert Dale, Editor

Computational Linguistics

Abstracting and indexing
ACM Computing Reviews
Annual Review of Information Science and Technology
Arts and Humanities Citation Index
CompuMath Citation Index
Compuscience Database
Computer Abstracts (Emerald Abstracts)
Computer and Control Abstracts (INSPEC Section C)
Computer and Information Systems Abstracts
Computer Literature Index
Current Contents/Arts and Humanities
Current Contents/Engineering Computing Index
EI Compendex
Engineering Index Monthly
Journals Citation Reports/Social Sciences Edition
Linguistic Bibliography
Linguistics Abstracts
Linguistics and Language Behavior Abstracts
Mathematical Reviews/MathSciNet
MLA International Bibliography
Research Alert
Science Citation Index
Science Citation Index Expanded
Social Sciences Citation Index/Social SciSearch
SoftBase

Computational Linguistics is the longest-running publication devoted exclusively to the design and analysis of natural language processing systems. From this highly-regarded quarterly, university and industry linguists, computational linguists, artificial intelligence (AI) investigators, cognitive scientists, speech specialists, and philosophers get information about computational aspects of research on language, linguistics, and the psychology of language processing and performance.

Volume 34 Issue 1 Highlights

MIT Press Journals | 238 Main Street, Suite 500 | Cambridge, MA 02142 USA
Tel: 617-253-2889 | US/Canada: 800-207-8354 | Fax: 617-577-1545
http://mitpressjournals.org/compling

ISSN 0891-2017 | E-ISSN 1530-9312