



Corpus Linguistics 25 Years on

Edited by
Roberta Facchinetti

Corpus Linguistics 25 Years on

LANGUAGE AND COMPUTERS:
STUDIES IN PRACTICAL LINGUISTICS

No 62

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

Corpus Linguistics 25 Years on

Edited by
Roberta Facchinetti



Amsterdam - New York, NY 2007

Cover design: Pier Post

Online access is included in print subscriptions:
see www.rodopi.nl

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-13: 978-90-420-2195-2

©Editions Rodopi B.V., Amsterdam - New York, NY 2007

Printed in The Netherlands

Contents

Introduction <i>Roberta Facchinetti</i>	1
1. Overviewing 25 years of corpus linguistic studies	
Corpus linguistics 25+ years on <i>Jan Svartvik</i>	11
Corpus development 25 years on: from super-corpus to cyber-corpus <i>Antoinette Renouf</i>	27
Seeing through multilingual corpora <i>Stig Johansson</i>	51
Corpora and spoken discourse <i>Anne Wichmann</i>	73
2. Descriptive studies in English syntax and semantics	
An example of frequent English phraseology: distributions, structures and functions <i>Michael Stubbs</i>	89
The semantic properties of <i>going to</i> : distribution patterns in four subcorpora of the <i>British National Corpus</i> <i>Ylva Berglund and Christopher Williams</i>	107
The superlative in spoken English <i>Claudia Claridge</i>	121
Semantically-based queries with a joint <i>BNC/WordNet</i> database <i>Mark Davies</i>	149
Size matters – or <i>thus can meaningful structures be revealed in large corpora</i> <i>Solveig Granath</i>	169
Inversion in modern written English: syntactic complexity, information status and the creative writer <i>Rolf Kreyer</i>	187

The filling in the sandwich: internal modification of idioms <i>David C. Minugh</i>	205
NP-internal functions and extended uses of the ‘type’ nouns <i>kind, sort, and type: towards a comprehensive, corpus-based description</i> <i>Liesbeth De Smedt, Lieselotte Brems and Kristin Davids</i>	225
3. Second Language Acquisition, parallel corpora and specialist corpora	
Student writing of research articles in a foreign language: metacognition and corpora <i>Francesca Bianchi and Roberto Pazzaglia</i>	259
The structure of corpora in SLA research <i>Ron Cowan and Michael Leiser</i>	289
The path from learner corpus analysis to language pedagogy: some neglected issues <i>Nadja Nesselhauf</i>	305
Exploiting the <i>Corpus of East-African English</i> <i>Josef Schmied</i>	317
Transitive verb plus reflexive pronoun/personal pronoun patterns in English and Japanese: using a Japanese-English parallel corpus <i>Makoto Shimizu and Masaki Murata</i>	333
The retrieval of false anglicisms in newspaper texts <i>Cristiano Furiassi and Knut Hofland</i>	347
Lexical semantics for software requirements engineering – a corpus-based approach <i>Kerstin Lindmark, Johan Natt och Dag, Caroline Willners</i>	365

“In studying corpora we observe a stream of creative energy that is awesome in its wide applicability, its subtlety and its flexibility.”

(John Sinclair, Introduction to *How to Use Corpora in Language Teaching*, Amsterdam: Benjamins, 2004, p.1)

“I CAME, but where are we going?”

(Title provided by Jeffrey Leech for the Panel of ICAME-25)

Introduction

The Conference that took place at the University of Verona on 19-23 May 2004 was intended to mark the silver jubilee of the meetings and conferences of the International Archive of Modern and Medieval English (ICAME) Association. As such, it attracted a number of scholars from all over the world, who contributed over a hundred paper presentations, thus leading the organisers to set up four parallel sessions, which then materialised as two separate books of proceedings; the first one – co-edited by the present author together with Matti Rissanen – was published early this year under the title *Corpus-based Studies of Diachronic English* (Bern, Peter Lang), as indeed it focuses on diachronic studies, while the present volume is related to synchronic research and aims to provide a fairly broad and thematic overview of the work undertaken in the field of computerised corpus linguistic studies from their origin to the present day, without overlooking their future prospects. The papers are grouped under three separate headings, which will be illustrated in detail in the following sections.

Section 1: Overviewing twenty-five years of corpus linguistic studies

Looking back “to the stone age of corpus linguistics” (p. 12), the time of language corpora B.C., that is Before Computers, **Jan Svartvik** opens the lively carousel of detailed reports delivered by the plenary speakers of the conference, by recalling the massive corpus-based work carried out by James Murray, Alexander Ellis, Otto Jespersen, and Charles Fries. When the era of computerised corpora started, from the sixties onwards, two linguistic ‘households’ appeared and confronted each other: ‘armchair linguists’ and ‘corpus linguists’. With great verve, Svartvik illustrates these two eras, how he has lived through them, how he has decided to become and to remain a corpus linguist, and finally how he has contributed to the realisation of the first family of corpora – the *Brown* corpus and its cognates.

Antoinette Renouf complements Jan Svartvik’s paper perfectly and widens its breadth, by providing a plethora of information and data concerning corpus development as it used to be in the early years of corpus linguistics, as it is nowadays and as it might be in the future. With detailed exemplifications, she touches on a number of corpora, some of which she has helped create herself: the

Birmingham Corpus, the *Bank of English*, the *British National Corpus*, open-ended corpora of journalism, and even the World Wide Web used as a corpus. She discusses the major motivating sources underpinning the development of these and other corpora and touches on theoretical issues like corpus ‘representativeness’ and on terminological aspects like the difference between ‘corpus’, ‘database’ and ‘text archive’ or again between ‘historical’ and ‘diachronic’.

In “Seeing through multilingual corpora”, **Stig Johansson** opens up a new folder in the corpus linguistics archive and testifies to the keyness of multilingual/parallel corpora in linguistic research; indeed, by covering two or more languages, they enable us to see more clearly not only the characteristics of each language, but also the features shared by individual languages and, maybe, even what typifies Language in general. While illustrating the typicality of these corpora, Johansson also supplies readers with essential theoretical and terminological concepts, including the definition of ‘parallel’ and ‘translation’ corpus, the notions of ‘translation paradigm’, of ‘mutual correspondence’ between particular forms or categories of forms, of ‘zero correspondence’, in the case of no formal cross-linguistic correspondence, and finally of ‘parallel translation’. These concepts are exemplified mostly from the *English-Norwegian Parallel Corpus*, from the *Oslo Multilingual Corpus*, including German, English and Norwegian, and from the *English-Swedish Parallel Corpus*, all of the three owing a lot to Stig Johansson’s enlightened ideas.

Finally, from written to spoken: **Anne Wichmann** devotes her contribution to “Corpora and spoken discourse”, as the title reads; with richness of detail and well-grounded argumentation, she ponders the potential and limitations of computerised corpora of spoken English from their early stages to the present. She starts from the *London-Lund Corpus* and the *Spoken English Corpus*, the only two prosodically transcribed, then focuses on the *Cobuild Corpus* and on the *British National Corpus*, and finally deals with the specialised *Reading Emotion Corpus* and the *Oxford/Cambridge IViE Corpus*, whose software provides annotation tiers linked to the speech signal. Her overview ends up with the *Aix-Marsec Corpus*: its compilers exploited a dictionary to capture phonemic representation automatically. While discussing the different ways of identifying features of speech, Wichmann convincingly argues that transcription details may provide a unique source of contextual information and are thus important, if not mandatory, for proper corpus analysis, in terms of tagging, parsing, and demographic/situational details. Yet she also warns that the analysis of transcriptions should not mislead the user into thinking the spoken recordings themselves are no longer relevant.

Section 2: Descriptive studies in English syntax and semantics

The description of language and the exploration of its structures, with special perspective on the interconnections between syntax and semantics, have always been among the privileged areas of study in corpus linguistics, if not the most privileged. Indeed, it is now widely acknowledged that grammar and lexis,

structure and meaning, syntax and semantics are so interpenetrating that we cannot study one without at least casting an eye on the other. The variety of topics raised in this section bear witness to the tremendous development which has unfolded in this field thanks to corpus linguistics.

Perhaps the most telling illustration of how strongly interrelated grammar and lexis are is provided by **Michael Stubbs'** plenary paper on Phraseology. To illustrate the different applications of linguistic corpora in this field, first he sets some basic terminological pillars, like the definitions of '*n*-grams', 'phrase-frames', and 'PoS-grams'. Then, he provides examples of how to extract 4- and 5-grams from corpora like the *BNC*; thirdly, he illustrates interesting results concerning these idiomatic form-meaning complexes, like the fact that most of them have specialised pragmatic, evaluative connotations, primarily pertaining to information-managing and text-structuring. Finally, while highlighting the usefulness of automated systems of extraction of *n*-grams, Stubbs does not refrain from warning about some of their drawbacks and about the need to downsize – for the sake of adequate phraseological analysis – huge amounts of quantitative data extracted from large corpora like the *BNC*.

The size of a corpus is actually among the key topics of **Solveig Granath's** paper, entitled "*Size matters – or thus can meaningful structures be revealed in large corpora*". She begins her analysis by examining small-sized corpora such as those of the *Brown* family, and notices that sentences featuring inversion and beginning with *thus* exhibit four alternative ways of ordering the subject and the verb – namely Subject-Verb, Auxiliary-Subject-Verb, Auxiliary-Verb-Subject, and Verb-Subject. Overall, the number of these sentences in the *Brown* family of corpora appears to be rather low. Then she checks these data against much larger corpora comprising issues of *The Guardian* and *The Observer* on CD-ROM; here, her results suggest that inversion is much less marginal than indicated in the first part of the research. However, both the standard million-word corpora and the larger British newspaper corpora show that inversion is used after sentence-initial *thus* not only when the verb phrase consists of a single verb but also when it includes an auxiliary. Interestingly, these alternatives are not mentioned in reference books. Indeed, on the one hand, Granath testifies to the usefulness of both small-sized and large-sized corpora and, on the other, convincingly complements and even corrects grammar books.

Inversion is also the topic of **Rolf Kreyer's** paper, in which he provides a discourse-functional, corpus-based account of the construction at issue, thus combining the strengths of the human researcher and of the computer. To do so, he analyses the 'Written-academic' and 'Prose-fiction' genres of the *BNC* and investigates the relationship between three factors which may affect the occurrence of inversion, namely (a) syntactic complexity, (b) information status, and (c) text-structuring and immediate-observer-effect. More specifically, he studies the distribution of syntactic complexity and information status across preposed and postposed constituents. The analysis and discussion of his results lead the author to conclude that such data can be explained as deriving from two superordinate discourse functions: inversions may serve either text-structuring

purposes or may be used to simulate natural perception, thus creating the illusion on the part of the reader to be an immediate observer of a particular scene or event.

Still within the context of lexico-grammar, **Liesbeth De Smedt, Lieselotte Brems** and **Kristin Davidse** (“NP-internal functions and extended uses of the ‘type’ nouns *kind*, *sort*, and *type*: towards a comprehensive, corpus-based description”) provide a grammatical description and a systematic classification of the uses of ‘type’ nouns. After reviewing the theoretical categorisations provided by other scholars, they set out their own descriptive framework, accounting for the different functional configurations found in their data, which are drawn from *The Times* subcorpus of the *COBUILD* corpus and from the *COLT* corpus. Their results lead them to argue that six functionally distinct categories of ‘type’ noun uses can be distinguished on account of semantic and formal criteria. Particular attention is paid to the delineation and elucidation of the NP-internal functions of ‘type’ nouns – head, modifier and postdeterminer uses – which have previously been relatively neglected in the literature.

Moving from syntactic patterns to semantic values, **Ylva Berglund and Christopher Williams** focus on the meanings of the semi-modal *be going to* and, like Stubbs and Granath before them, also take up the issue of the size of corpora, by highlighting (a) the risks of generalising small figures, which may lead to arbitrary conclusions, and (b) the importance of basing one’s studies on a wide spectrum of data. Yet interesting results can still be drawn from relatively small corpora, which may serve as good, promising starting points; indeed, by making use also of *BNC Baby*, the authors show that the Academic, Fiction and Spoken genres behave similarly with reference to *be going to*, exhibiting some prevalence of intentional values, while the News genre favours Prediction. In their analysis, they also discuss the correlations between subject number and semantic values, bearing in mind that context plays a fundamental role in determining meaning and that several concomitant aspects have to be considered in order to reach relatively conclusive results.

Claudia Claridge perfectly exemplifies such importance of the role of syntactic contexts for the qualification of meaning, by focusing on the interconnections between the semantic, functional, morphological and syntactic matters involved in the superlative forms, be they inflectional or periphrastic. To do so, like most of the contributors to this section, she also makes use of the *BNC*, specifically its spoken demographic part, for a total of ca. 4 million words. Her data testify to the fact that the pattern under scrutiny is relatively rare and consequently does not appear to be a prominent way of marking high degree in conversation. The author also illustrates that a very important purpose of the superlative is not primarily comparison but rather evaluation and intensification; hence, it carries a high potential for uses other than factual comparison in everyday conversation. This and other aspects put forward in her exhaustive paper prove that the superlative needs to be interpreted as an important device exploited in involved, emotive language styles.

The lexicon is next as the focus of **David Minugh**'s paper entitled "The filling in the sandwich: internal modification of idioms", analysing a series of relatively well-known idiomatic expressions and examining the extent to which these clusters can be internally expanded so as to link them into the discourse where they are used. To obtain a broad sample from various types of English, the author exploits a number of corpora, mostly journalistic; then, he selects a total of 55 idioms from the *Collins Cobuild Dictionary of Idioms* and searches for them within his corpus. His data allow him to classify the idioms according to a five-category system: (a) simplex, (b) modified, (c) names, (d) exemplifications, and (e) anchorings. He focuses particularly on anchorings, where the idiom is not merely modified, but actually hooked into ('anchored in') the discourse by means of adjectives or other words. Interestingly, despite the claims from previous scholars, in Minugh's corpus anchorings appear to be just a minor option, rather than a major pattern. Among the most interesting findings of his in-depth analysis is the fact that some of the idioms are truly self-contained, hardly admitting any modification at all, while others intrinsically demand modification or exemplification.

Finally, going beyond the values of one single lexical pattern so as to identify a systematic semantic net of values among different lexemes, **Mark Davies** illustrates the advantages of creating a joint database of *BNC* and *WordNet*, so as to use simultaneously both the frequency and collocational information provided by the *BNC* and the extensive semantic hierarchies encoded in *WordNet*. The author exemplifies some basic queries via this interface and shows how they are carried out from the point of view of the end user and also how they are processed in terms of the underlying *WordNet* and *BNC* databases. Hence, for example, it is possible to identify all the collocations involving a synonym of an adjective followed by a noun or to discover strings of synonyms expressing a given semantic concept; or again to select lists of hyponyms and hypernyms; to trace meronyms and holonyms (larger units a given word is part of); and finally to find which words occur with greater or lesser frequency in different registers. Alongside these doubtless positive aspects, however, the author does not downplay some problematic issues that still need to be solved, determined partly by the fact that the database tends to overgenerate results, and partly by the fact that there may be mismatches, since the database is based primarily on American English, while the *BNC* is exclusively British English data.

Section 3: Second Language Acquisition, parallel corpora and specialist corpora

The final section in the book covers seven papers focussing on different, but strongly interrelated fields; indeed, almost all of them touch on issues referring to the value of corpora in Second Language Acquisition, both in terms of theoretical discussion and of practical applications. So, **Ron Cowan and Michael Leeser** exploit a self-made corpus of Spanish learners of English to demonstrate how large-scale corpora of adult second language learner errors could help answer

specific research questions in SLA. The authors first illustrate the basic features of L2 learner corpora and then identify and track a few errors of Spanish learners in written English, focussing on (a) which errors persist as the students progress towards stabilised grammars and (b) which errors disappear as a result of continued exposure to the target language and instruction. Cowan and Leeser conclude that their results already allow us to draw some interesting conclusions about the structure of L2 learner corpora for SLA research, but they still need to be tested against a wider corpus of analysis.

Working against the background of German-speaking learners of English, **Nadja Nesselhauf** attempts to shed light on the nature and extent of the difficulties of L2 learners, particularly with reference to collocations. Her paper – entitled “The path from learner corpus analysis to language pedagogy: some neglected issues” – confirms that the results from learner corpus studies can contribute to the improvement of (foreign) language teaching. More specifically, she argues that the path from learner corpus analysis to language pedagogy is not as direct as has sometimes been assumed. She also points to some directions of study in this field, by focusing particularly on how features of language can be selected for teaching on the basis of results from learner corpus studies.

Still within the context of SLA, though with a more direct focus on English for Specific Purposes, **Francesca Bianchi and Roberto Pazzaglia** report on the creation and use of a corpus of Psychology research articles in a module designed to teach Italian undergraduates how to write research articles in English. The authors show that very simple tools – like a home-made non-tagged corpus, concordance lines, clusters, wordlists, and keyword lists – can be of great use for this purpose; however, the corpus alone does not suffice; indeed, students need to be made aware of conventions pertaining to specific disciplines and of the advantages offered by complying with them, in order to be guided in the discovery of the moves and steps that characterise scientific papers.

Working in the border zone between second language acquisition and language variation, **Josef Schmied** provides a detailed overview of the current status and future plans for the *Corpus of East African English*. His study mainly concentrates on the methodological problems related to exploitation tools, starting from the individual text through the entire corpus. Moving from the widely acknowledged view that variation across text-types is more pronounced than variation across varieties, he convincingly argues for a cline of cultural differences in ESL varieties, which can be summarised as follows: (a) both the lexicon and pragmatics are clearly marked by cultural differentiation; (b) in contrast, grammar is only rarely an indicator of cultural variation, although the use of modal verbs or of passives illustrates subtle differences here as well; (c) finally, idiomaticity is also an area where users are unaware of culture-specific influences, although metaphors and special collocations very often occur.

Shifting from the analysis of one language to the structural differences between languages, **Makoto Shimizu and Masaki Murata** (“Transitive verb plus reflexive pronoun/personal pronoun patterns in English and Japanese: using a Japanese-English parallel corpus”) illustrate a few English Japanese parallel

corpora and then focus on one of them, the *Context Sensitive and Tagged Parallel Corpus*. Their aim is to extract examples of the English patterns ‘transitive verb followed by reflexive/personal pronoun’ (Vt SELF) automatically together with their counterparts in Japanese, and manually analyse them syntactically and semantically, so as to highlight which types in one language correspond to which types in the other language. Their data point to five strategies to make Japanese expressions correspond to the type Vt SELF in English. Indeed, even though the choice of a strategy is at the translator’s discretion, it seems that there exist some tendencies, like (a) to intransitivise the transitive verb, namely just delete the reflexive pronoun; (b) to paraphrase the verb, often with bound morphemes; (c) or to change the reflexive pronoun into a noun.

Still within the cross-linguistic perspective, **Cristiano Furiassi and Knut Hofland** illustrate a project aimed at building a specialised corpus of Italian newspaper texts to be used for retrieving new false anglicisms and describe in great detail the compilation of such a corpus and the computational technique implemented in order to isolate these words and phrases. Besides offering a large searchable linguistic database, computer corpora prove to be very useful in this respect, since they enable the researcher (a) to evaluate whether a certain entry is to be considered a false anglicism or not and (b) its most typical context of occurrence. However, the authors acknowledge that the computational techniques employed still do not seem to be sufficient to handle the complex and many-faceted phenomenon of false anglicisms; indeed, along with automatic processing, a final manual scanning is indispensable.

The final paper of this section, which also closes the present volume, is by **Kerstin Lindmark, Johan Natt och Dag and Caroline Willners**, who focus on terminology for software requirements. They start their paper by posing a set of questions, among which are the following: (a) Is there a specific terminology for software requirements? (b) Does such terminology coincide with the terminology of other documentation within the domain? (c) Would it be possible to structure such a domain-specific vocabulary into *Wordnet* parts that could be of relevance? And finally, (d) will a *Wordnet*-type lexicon be advantageous for industrial analysis and grouping of software requirements? After illustrating the construction of a lexicon using a semi-automated, corpus-based approach, the authors remark that it is indeed possible to automatically extract domain-specific terms from a collection of natural language texts; it is also possible to establish smaller sub-nets for different fields within the domain; finally, automatic extraction of relations by way of lexico-syntactic patterns is shown to prove helpful for establishing semantic relations between concepts.

From terminological issues to practical applications, from theoretical research studies to applied linguistic problems, from monolingual to multilingual corpora and corpus compilation details: all the aspects investigated in this volume are an impressive testimony to the progress that has been achieved over twenty-five years of corpus compilation, development, and use. Undoubtedly, there are still gaps in the coverage and many uncharted areas, which certainly call for

further analysis and – I would like to add – for new conferences and meetings of the ICAME Association in the years to come.

Roberta Facchinetti

University of Verona
October 2006

Overviewing 25 years of corpus linguistic studies

This page intentionally left blank

Corpus linguistics 25+ years on

Jan Svartvik

Lund University

Abstract

In the history of English language research on computerised corpora, the year 1977 marks an important event with the birth of ICAME – the International Computer Archive of Modern and Medieval English – which set off international co-operation on a large scale. The use of computer corpora, from being a fringe activity, has become a mainstream methodology. Yet there was corpus life also before ICAME. I have sometimes been asked why, in the unsupportive linguistic environment of the 1960s, I chose to become ‘a corpus linguist’ – there might have been moments when being so named felt like discovering your name on the passenger list for the Titanic. This contribution is very much a personal memoir of those early days when the first corpora were being compiled, when computers were rare, expensive, unreliable and inaccessible to ordinary folk – huge machines located inside glass doors and operated by engineers dressed in white coats, and when CD only stood for Corps Diplomatique.

Treading on Roberta Facchinetti’s classic ground, I find it appropriate to open my talk on early corpus plumbing with a piece of poetry from the Prologue of William Shakespeare’s *Romeo and Juliet*:

Two households, both alike in dignity,
In fair Verona, where we lay our scene,
From ancient grudge break to new mutiny,
Where civil blood makes civil hands unclean.

Especially in the early days of corpus linguistics, there were two linguistic households – corpus linguistics and armchair linguistics. Their grudge has not (to my knowledge) led to bloodshed but, as Charles Fillmore put it:

Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable chair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, ‘Wow, what a neat fact!’, grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn’t anybody exactly like this, but there are some approximations).

Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations).

These two don't speak to each other very often, but when they do, the corpus linguist says to the armchair linguist, 'Why should I think that what you tell me is true?', and the armchair linguist says to the corpus linguist, 'Why should I think that what you tell me is interesting?' (Fillmore 1992: 35)

Today, I believe, such questions are more rarely put than in the 1960s. Those of us who have been on the corpus bandwagon since it was a little donkey cart – when, as Geoffrey Leech says (1991: 25), “corpus work was, indeed, little else but donkey work” – find that the use of computer corpora, from being a fringe activity, has become a mainstream methodology, attracting more and more linguists. It is of course only circumstantial evidence, but a Google check gave 12,400 hits for the collocation *transformational grammar*, 11,600 hits for Chomsky's latest *minimalist program*, as compared with 34,500 hits for *corpus linguistics*.

Italians have a strong sense of history, which probably explains why Roberta asked me to talk about 'Corpus linguistics 25+ years on'. Yes, there was indeed corpus life before ICAME, and I take my mission at this conference to go further back than the beginnings of ICAME – back to the stone age of corpus linguistics. Neolithic food is popular these days, and Neolithic corpus linguistics may be of some interest to the young corpus-driven linguists who have gathered here in Verona for this 25th ICAME meeting. If you lean back in your comfy chairs I'll take you back to a time when there were no personal computers, no web, no e-mail, no mobile phones, no Google, and no electronic corpora.

While it is natural today to take 'corpus linguistics' to mean 'electronic corpus linguistics', we must not forget that there were language corpora BC, i.e. 'before computers'. There were lexicographical corpora, especially that used for the *Oxford English Dictionary* which, by the time the first edition was completed in 1928, had accumulated over four million citation slips from two thousand volunteer readers. Not having access to a computer, James Murray solved the problem by employing his family members – much of the work of alphabetising and sorting the slips was done by Murray's many children (1977: 178-179).

Before BC there were also dialect studies, such as *The Existing Phonology of English Dialects* by Alexander J. Ellis, for which he obtained information and assistance from over 800 persons. The early 20th-century grammarians, especially

the great Dane Otto Jespersen, made extensive use of authentic texts. Jespersen wrote:

It is impossible for me to put even a remotely accurate number on the quantity of slips I have had or still have: a lot of them have been printed in my books, particularly the four volumes of *Modern English Grammar*, but at least just as many were scrapped when the books were being drafted, and I still have a considerable number of drawers filled with unused material. I think a total of 3-400,000 will hardly be an exaggeration. (Jespersen 1938: 213-215; translation by D. Stoner)

Another pioneer in corpus linguistics BC was Charles Carpenter Fries at Ann Arbor, Michigan. *The Structure of English* is an analysis of

a large body of actual English speech observed and recorded in a university community in the North-Central part of the United States [...] The materials which furnished the linguistic evidence for the analysis and discussions of the book were primarily some fifty hours of mechanically recorded conversations on a great range of topics – conversations in which the participants were entirely unaware that their speech was being recorded. (Fries 1951: 3)

In the 1960s universities began to set up computer units, mainly to be used by scientists – indeed the first mainframes were referred to as ‘mathematics machines’. In those days computers were rare, expensive, unreliable and inaccessible to ordinary folk – huge machines located inside glass doors and operated by engineers dressed in white coats.

Here is my story – how it happened that I became involved in the study of English by computer. In the late 1950s, a postgraduate student of English at the University of Uppsala in Sweden, I went through the mandatory series of courses, chiefly aimed at the historical study of the language. Then came the time to begin thinking about a possible subject for a thesis. I naturally chose the language option, but in those days theses usually dealt with such topics as the history of the language, place names, dialectology, 16th-century orthography and pronunciation, and the philological study of old texts. Proficiency in the modern language, especially documented in translation, was of course important in undergraduate studies, but hardly anyone was doing research in Modern English. Serendipitously, as I was one day browsing through the departmental library, my eyes fell on an article in the latest issue of *English Studies* called ‘Relative clauses in educated spoken English’ by a certain R. Quirk (1957). I was instantly hooked by this approach: studying English that was contemporary and spoken, based on a corpus of audio recordings, both surreptitious and non-surreptitious – it even smelled of cloak and dagger. Armed with a British Council scholarship I left for north-east England to spend the 1959-1960 academic session at the University of Durham under the guidance of R. Quirk.

Leaving Durham with the embryo of a thesis to be completed in Uppsala, I was decorated with a licentiate degree in the spring of 1961. A week after submitting my thesis, I received a letter from Randolph Quirk offering me an assistantship on the Survey of English Usage at University College London where he had moved from Durham, an offer which I naturally could not refuse.

In those days, Uppsala was hardly a hotbed of linguistics, but one thrilling moment was a lecture in 1953 to a packed auditorium by Michael Ventris about the success of decipherment of Linear B undertaken in collaboration with John Chadwick at Cambridge University – an outstanding and revolutionary scientific achievement. The transfer to London meant a terrific uplift for me. It turned out to be a four-year period at the Survey of English Usage, which gave me the opportunity of working also with many young postgraduate students attached to the English Department, and now familiar names in English linguistics, including Sidney Greenbaum, Geoffrey Leech, and David Crystal. Upstairs from the Survey Michael Halliday had set up a linguistics department with colleagues like Bob Dixon, Rodney Huddleston and Dick Hudson, and next door was the Phonetics Department with A. C. Gimson, J. D. O'Connor and John Wells. Just down the road at the School of Oriental and African Languages, R. H. Robins lectured on general linguistics. We could widen our horizons still further by attending meetings of the London University Linguistics Circle at Bedford College. I particularly remember a talk there by John Lyons who, returning from a period at Indiana University, declared he was 'a transformational linguist' – a bold confession to make to a largely Firth-inspired audience.

Our first computer project on the Survey was based on some of the early collected spoken texts and submitted to the Ninth International Congress of Linguists, in Cambridge, Massachusetts (Quirk et al. 1964). It was a study of the correspondence of prosodic to grammatical features in English, a topic to which I was later to return. This research also gave me my first idea of how a computer's ability to store and manage information could be put to good use for a linguist.

One day, I believe it was in 1963, W. Nelson Francis from Brown University turned up at University College, walked into the office and dumped one of those huge computer tapes on Randolph Quirk's desk with the accompanying words "Habeas corpus". This was the *Brown Corpus* which W. Nelson Francis and Henry Kučera had just completed – the first large computerised text collection of American English for linguistic analysis. Its size of one million words and principles of text selection set a pattern for the creation of new corpora and for the development of tools to analyse these corpora. Over the years, this corpus has been used in English departments all over the world as a source of data or a means of exploring the ways in which computers can be employed in language research.

One of the milestones in the history of 20th-century linguistics was of course the publication in 1957 of *Syntactic structures* by Noam Chomsky, which imposed a constricting religiosity upon America and set off, for decades, the dominance of worldwide transformational generative linguistics. His view on the

inadequacy of corpora, and adequacy of intuition, became the orthodoxy of succeeding generations of theoretical linguists:

Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. (Chomsky 1958: 159)

Still, while the Survey staff certainly tried to learn about the new transatlantic ideas, Chomsky's theory had little effect on our daily work.

I have sometimes been asked why, in the unsupportive linguistic environment at the time, I chose to become 'a corpus linguist' – there might have been moments when being so named felt like discovering your name on the passenger list for the *Titanic*. One reason is that, in the long Scandinavian philological tradition in English studies, the text was central. Another reason is of course that to a non-native speaker of the language, the armchair approach of introspection is effectively ruled out. This may help to explain why certain parts of the world outside the English-speaking countries, such as northern Europe, were early strongholds of corpus linguistics. Before ICAME, corpus linguistics was buttered thin geographically. But a look at the list of participants at this ICAME conference indicates that corpus linguistics has now gone global.

Still, the word 'corpus' was not a common term in the early days of the Survey of English Usage. In his plan for the Survey (1960), Randolph Quirk talks about a 'Descriptive Register', 'primary material' and 'texts'. I recall one discussion over morning coffee in the UCL common room about whether the correct plural of *corpus* should be *corpuses* or *corpora*. The session came to an abrupt stop when somebody suggested: 'I think it's *corpi*'.

At least in the Anglo-Saxon world, questions of usage were (and probably still are) largely answered by notions of appropriate usage as entertained by, among others, the Fowler brothers rather than professional statements based on actual, documented usage. While there is considerable public interest in questions of usage, it seems hard work for linguists to convince the public of the validity of their advice, even when supported by actual usage, and to bring home the notion that grammar is not synonymous with 'linguistic etiquette'. Also, many of the practitioners who give advice on usage lack both linguistic competence and real data, as Dwight Bolinger has pointed out:

In language there are no licensed practitioners, but the woods are full of midwives, herbalists, colonic irrigationists, bonesetters, and general-purpose witch doctors, some abysmally ignorant, others with a rich fund of practical knowledge... They require our attention not only because they fill a lack but because they are almost the only people who make the news when language begins to cause trouble and someone must answer the cry for help. Sometimes their advice is

sound. Sometimes it is worthless, but still it is sought because no one knows where else to turn. We are living in an African village and Albert Schweitzer has not arrived yet. (Bolinger 1980: 1)

The goal of the Survey of English Usage was in fact to describe the grammatical repertoire of adult educated native speakers of British English: their linguistic activity, ranging from writing love letters or scientific lectures to speaking upon a public rostrum or in the relaxed atmosphere of a private dinner party. Since native speakers include lawyers, journalists, gynaecologists, school teachers, engineers, and a host of other specialists, it follows that no grammarian can describe adequately the grammatical and stylistic properties of the whole repertoire from his own unsupplemented resources: introspection as a sole guide star is clearly ruled out (Quirk 1960).

Like the *Brown Corpus* for American English, the *Survey of English Usage Corpus* for British English was to total one million words collected from a planned programme of varieties, but unlike Brown it was to include both spoken and written material. There was no question of attempting to match proportions with the statistical distribution of the varieties according to normal use: this would obviously have obliged us to assign over 99 per cent of the corpus to the preponderant variety of conversation between people on an intimate footing. Instead, we saw the overwhelming criterion as being the relative amount of data that would be required to represent the grammatical/stylistic potential of a given variety. The most difficult and time-consuming part of the work was of course the transcription of the audio recordings, especially those of spontaneous, interactive speech. Since we believe that prosody is a part of grammar, the decision was taken to include a transcription which was sensitive to a wide range of prosodic and paralinguistic features in the spoken realisation as heard on the recording. This system was documented by David Crystal and Randolph Quirk (1964) and further elaborated by Crystal in his PhD thesis (1969).

It was of course never envisaged that any corpus, necessarily finite (at least not in those pre-web days) would itself be adequate for a comprehensive description of English grammar. From the outset, elicitation tests with native subjects were envisaged as an essential tool for enlarging upon corpus-derived information and for investigating features perhaps not found in the corpus at all.

While we did not buy Chomsky's ideas wholesale, they nevertheless inspired us to undertake some related research. One basic concept in his theory was grammaticality: "the fundamental aim" of a grammar, he wrote in *Syntactic structures* (1957), is to account for "all and only the grammatical sentences of a language". To us on the Survey, surrounded by masses of authentic language data, both spoken and written, drawing the line between grammatical and ungrammatical sentences seemed a big problem. You will realise that our detailed analysis of a corpus consisting of real-life language was very much swimming against the tide of the mainstream Chomskian view of language!

We undertook informant-based studies trying to devise a technique for establishing degrees and kinds of acceptability of different English sentences

(Quirk and Svartvik 1966). We had found that direct questioning (such as ‘Is this a grammatical sentence?’) was the least reliable technique. Our way of improving on the direct question technique (which we called ‘judgement test’) was to present informants with sentences on which they were required to carry out one of several operations (hence called ‘operation tests’) which were easy to understand and to perform. An example would be to turn the verb in the present tense into the past tense, and it would be left to informants to make any consequential changes that then seemed necessary. An example: when asked to turn the verb in *They don’t want some cake* into the past tense, 24 of the 76 informants replaced *some* with *any*, and several others showed obvious discomfort over *some*, with hesitations and deletions.

The results indicated that clear-cut categorisation is futile and may actually inhibit our understanding of the nature of linguistic acceptability. In fact, the judgements of informant groups occurred anywhere throughout the entire range between unanimous acceptance and unanimous rejection. Testing acceptability in this way is not basic corpus linguistics but rather an extension of it, so as to investigate not only linguistic performance but also linguistic attitudes, and both techniques were part of the original Survey plan.

Most of my work on the Survey consisted of grammatical analysis of texts by marking paper slips which were stored in various categories in filing cabinets. In those days computers were, as I said, rare, expensive and unreliable. In the company of Henry Carvell, a Cambridge mathematician turned programmer on joining the Survey, I spent many late nights in Gordon Square to get inexpensive off-peak access to the Atlas machine, programmed by punched paper tape. The results of our research on linguistic classification were published in Carvell and Svartvik (1969). The topic was the pursuit of suitable computational methods of analysing real-language data, where we used a program primarily intended for the classification of bacteria – looking in the rear mirror, a pretty bold undertaking, considering the negative attitude to taxonomy in the dominant linguistic climate at the time. We found the concept of gradience to be true in the material: the results gave us a few fairly distinct classes with some partially overlapping classes.

Also the topic of my PhD thesis, *On voice in the English verb* (1966), can be said to have been inspired by TG. In his early theory, Chomsky derived passive sentences from kernel active sentences – the prime illustration of TG, at least at the time – claiming that every active sentence with a transitive verb can be transformed into a corresponding passive sentence. The idea of representing the active-passive relation in terms of transformations was not new – Jespersen talked about the ‘turning’ and Poutsma of the ‘conversion’ of the verb form from one voice to another, but it was only in transformational theory that the use of transformation was extended, formalised and systematically incorporated into a unified grammatical framework. The validity of the huge claim for active-passive transformation seemed worth investigating. After all, a pair of active-passive sentences like *I play football* and *Football is played by me* is enough to make any native speaker dubious. Yet considerations other than linguistic may influence

informant judgements: in one informant test with English students, who were asked to mark, in an answer sheet, the acceptability of the active-passive pair *I have a black Bentley* and *A black Bentley is had by me*, one student, in addition to rejecting the passive submitted to him by a rather threadbare foreign postgraduate student, wrote this marginal comment: “And I don’t think you can afford a black Bentley in the first place” – which was of course a correct observation. In fact, asked to estimate its linguistic deviance, 73 informants rejected, two queried, and one accepted the sentence *A black Bentley is had by me*. On the other hand, 74 informants accepted, one queried, and one rejected the active analogue *I have a black Bentley*. In a footnote of my thesis I wrote: “A much more acceptable passive of *have* than *A nice little car is had by me ... is A good time was had by all ...* Although this sentence appears to contradict the general rule that *have* in the sense of ‘own, enjoy’ does not occur in the passive, it is, rather, the exception that proves the rule. In fact, it is precisely because of its deviance (and institutionalisation) that this sentence achieves its special effect of facetiousness” (1966: 165).

Over 300,000 words in some coexisting varieties of present-day English, spoken and written, were subjected to a variety of analyses which indicated that syntactic relationships can, or should, be expected to be multidimensional rather than binary and, in order to find this network of relations, it was best to cast the net wide. Some other results were these: the low proportion of passives with *by*-agents, about one in five; the restrictions on complex verb phrases in the passive as compared with the active, for instance the active *These Conservatives have not been winning seats lately* and the passive *Seats have not been being won by these Conservatives lately*; and the large variations in the frequencies of passives in different text categories. The conclusions state that there is in fact ‘a passive scale’ with a number of passive clause classes that have different affinities with each other and with actives, including both transformational and serial relations. For a corpus linguist it has been comforting to take in Michael Halliday’s support (most recently in Halliday 2005) by stressing the place of the corpus as a theoretical construct – to him data gathering from a corpus, especially a spoken corpus, and theorising are no longer separate activities, but the textual instance is valued as a window on to the linguistic system.

The thesis was well received at the University of Uppsala, my Swedish alma mater. Still, the book was essentially a British product, conceived in Durham and largely written in the exciting atmosphere of the Survey of English Usage at University College London, with several colleagues offering compelling stimulus and acute criticism. In Scandinavia, doctoral disputations are public (and, at the time, were performed in tails). They could be brutal in that the appointed critics might consider themselves future rivals for a chair and make a point of sinking their *respondents* into the slough of academic despond. Fortunately for me, the Uppsala Faculty wisely appointed as my opponent Nils Erik Enkvist, who was Vice-Chancellor of Åbo Akademi University, hence clearly with no axe to grind in future academic competition with me. In fact, the disputation act turned out to be one of the highlights of my life with a stimulating

discussion of my written endeavours, and a good time was had by all (as said by some people in the audience who had obviously read my thesis). After the event, my opponent confided to me that he had been troubled by the statistics in the page proofs I had sent him: lots of percentages were inaccurate. The reason was that I had worked them out on a slide rule, which of course gives only approximations! This is just one further point to show how much research conditions have improved since then.

Towards the end of the 1960s, Freeman Twaddell sent an invitation for me to be visiting associate professor in the Linguistics Department of Brown University. It was of course an offer I could not refuse in view of my previous positive experience of American academic life as an undergraduate at Duke University and the attraction of spending a year at the birthplace of the *Brown Corpus of American English* in the company of its compilers W. Nelson Francis and Henry Kučera.

While the *Brown Corpus* has been popular and used all over the world for over 40 years, it may not be inappropriate to remind latter-day corpus boffins that, in the sixties, there was unsupportive environment encountered in linguistics circles. Jan Aarts told me that, every year, he quoted to his students at the University of Nijmegen this anecdote in the words of W. Nelson Francis:

In 1962, when I was in the early stages of collecting the *Brown Standard Corpus of American English*, I met Professor Robert Lees at a linguistic conference. In response to his query about my current interests, I said that I had a grant from the U.S. Office of Education to compile a million-word corpus of present-day American English for computer use. He looked at me in amazement and asked, 'Why in the world are you doing that?' I said something about finding out the true facts about English grammar. I have never forgotten his reply: 'That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text.' (Francis 1982: 7-8)

But there was also unsupportive environment in other circles. Here is the opening passage of an after-dinner speech by W. Nelson Francis at the Windermere ICAME meeting in 1984:

You probably can't see it from where you sit, but some of you may have noticed that I am wearing a tie clip in the shape of a monkey wrench – or what I believe is called an 'adjustable spanner' in the curious dialect of this country. The story behind this peculiar piece of jewelry goes back to the early 60s, when I was assembling the notorious *Brown Corpus* and others were using computers to make concordances of William Butler Yeats and other poets. One of my colleagues, a specialist in modern Irish literature, was heard to remark

that anyone who would use a computer on good literature was nothing but a plumber. Some of my students responded by forming a linguistic plumber's union, the symbol of which was, of course, a monkey wrench. The husband of one of them, being a jewelry manufacturer, had a few of these clips made. I cannot say that they have become collectors' items, but I would certainly not part with mine.
(Francis 1985: 5)

The *Brown Corpus* was a pioneering project in being the first machine-readable corpus for linguistic research. At the time, processing large amounts of linguistic data required enormous computer resources. Henry Kučera mentioned that, when they produced the concordance of the corpus, they had to use the complete mainframe capacity of the Brown University Computer Unit – in all other university departments, computer-based research activities had to be suspended for a full day!

One joint project largely based on insights derived from research on the Survey was *A grammar of contemporary English* (Quirk et al. 1972) written by a foursome (by one reviewer referred to as 'the Gang of Four'). When work on this grammar began, all four collaborators were on the staff of the English Department, University College London. This association happily survived a dispersal which put considerable distances between us (at the extreme, the 5000 miles between Wisconsin and Europe). In those days with no personal computers, email and faxes – even electric typewriters were far from common – physical separation made collaboration arduous and time-consuming. Still, the book appeared in 1972. The original plan was to write a rather unpretentious undergraduate handbook but, with ambitions soaring, the printed book came to a total of 1120 pages. No wonder our Longman editor Peggy Drinkwater, equally professional and patient, used to refer to the ever growing manuscript as 'the pregnant brick'.

Our grammar was well received but, in the early eighties, we felt it was time to embark on an updated edition: this culmination of our joint work resulted in a grammar that was considerably larger and richer, *A comprehensive grammar of the English language* (Quirk et al. 1985). The grammarian is beset with a number of problems. One is the question of descriptive adequacy, as indicated in the openings lines of a review which appeared in *The Times* on 23 May 1985:

Writing a grammar of a living language is as muddy an undertaking as mapping a river. By the time you have finished, the rain has fallen, the water has moved on, the banks have crumbled, the silt has risen. With English having become the world language, in silt and spate with hundreds of different grammars, the project of making a comprehensive grammar of it is as Quixotic as trying to chart the Atlantic precisely.

It is typical of newspaper reviewers to focus on the changing language rather than general descriptive problems. It is not the grammar of English that has changed a lot – pronouns, for instance, have largely stayed the same for over four hundred years. The grammarian's real problem is choosing adequate descriptive categories and presenting them in an appropriate form of organisation.

Another problem for the grammarian is of course to find an audience for the book. *The Times* reviewer, Philip Howard, concludes by saying:

It is a prodigious undertaking. It is just a bit difficult to see who it is for, other than other writers of English grammars. You would be ill-advised to try to learn English from it.

Here Philip Howard was right. A pedagogical grammar has to be different from a descriptive grammar. Therefore we also wrote pedagogical grammars for advanced students (among them, Greenbaum and Quirk 1973, Leech and Svartvik 1975, Svartvik and Sager 1978).

After returning from London to Sweden, I was first a lecturer at the University of Göteborg, before being appointed in 1970 to be the incumbent of the Lund chair of English, which I sat on for the next 25 years. In the mid-seventies, the *London-Lund Corpus* was launched, with the aim of making the spoken part of the *Survey of English Usage Corpus* available in electronic form. Thanks to a generous grant from the Bank of Sweden Tercentenary Foundation, it was possible to employ a group of postgraduate students and get secretarial help to transfer the spoken material typed on paper slips in London to electronic form, a vast undertaking, considering the size of the material and the problem of finding ways of representing the detailed prosodic transcription in digital form. In 1980 we published a printed book, *A Corpus of English Conversation*, including 34 spoken texts from the magnetic tape (Svartvik and Quirk 1980). Later, the corpus became available on CD-ROM, and it is still one of the largest and most widely used corpora of spoken English, not least because it is prosodically annotated. The detailed annotation has facilitated numerous studies of lexis, grammar and, especially, discourse structure and function. Under the present director of the Survey of English Usage, Bas Aarts, the corpus has recently been enhanced by adding wordclass tags to it, using the ICE-GB scheme. In addition, the Survey has plans to digitise the original sound recordings to be supplied as a new resource.

Backtracking to the mid-seventies when the London-Lund project was launched, I had three main reasons for opting for research in spoken language. First, the *Brown Corpus* was a resource of machine-readable text exclusively for the medium of written language, and this was also the case with the on-going *Lancaster/Oslo-Bergen Corpus* project, which was the first of the *Brown Corpus* clones and designed to be a British counterpart of the American corpus. Furthermore, then available grammatical descriptions of English were almost exclusively based on written language. Yet the vast majority of English language use – perhaps 99 per cent – takes place in the spoken channel. Second, it seemed

a pity that the unique prosodic transcriptions of the Survey of English Usage should be restricted to the small number of research scholars who had physical access to the filing cabinets in London. Third, in the 1970s computers were becoming more widespread and efficient, opening up new exciting approaches to research in spoken English.

Today anybody anywhere in the world with a laptop, a CD unit and some off-the-shelf software can study different aspects of spoken English. The final product, the *London-Lund Corpus*, offered at cost to all interested colleagues in all parts of the world, was the result of research, recording, analysis and compilation extending over many years and involving a great number of colleagues on the Survey of English Usage at University College London and on the Survey of Spoken English at Lund University. But for the dedication and arduous teamwork by many students, there would be no *London-Lund Corpus*. Many other Lund colleagues contributed to the corpus and made extensive use of it – including Karin Aijmer (now professor at Göteborg University), Bengt Altenberg (later professor at Lund), Anna Brita Stenström (later professor at Bergen University), and Gunnel Tottie (later professor at the University of Zurich).

So, there was corpus life even before the birth of ICAME, which took place on an evening in February 1977 in Stig Johansson's Oslo kitchen. The founding fathers were Geoffrey Leech, who came from Lancaster with a suitcaseful of corpus texts, W. Nelson Francis, who was then guest professor at the University of Trondheim, Jostein Hauge, director of the Norwegian Computing Centre for the Humanities, Arthur O. Sandved, chairman of the English Department at the University of Oslo, Stig Johansson and myself. This was the beginning of international co-operation on a large scale. This group signed a document saying that the primary purposes of the International Computer Archive of Modern English will be:

- (1) collecting and distributing information on English language material available for computer processing;
- (2) collecting and distributing information on linguistic research completed or in progress on the material;
- (3) compiling an archive of corpuses to be located at the University of Bergen, from where copies of the material could be obtained at cost.

The last point is important: the generosity in making the *Brown Corpus* freely available for research provided a model for subsequent ICAME corpora.

In 1978 there appeared a modest newsletter called *ICAME News*, which later turned into the more prestigious *ICAME Journal*. The organisation has also organised annual conferences, among which this is the 25th. At first they were small affairs, a dozen or so people around a seminar table – no need for parallel sections then. I believe the Windermere ICAME meeting in 1984 attracted 48

people – a record at the time. It was indeed hard to find anybody interested in running this show – now it's become almost as attractive as organising the Olympic Games.

ICAME publications from the first ten years indicate frequent contributions from scholars at several universities, especially these: Birmingham, Lancaster and London in the UK; Amsterdam and Nijmegen in Holland; Liège and Brussels in Belgium; Helsinki in Finland; Bergen and Oslo in Norway; Uppsala, Stockholm, Göteborg and Lund in Sweden; Toronto in Canada; and Brown and Boston in the United States.

During its existence this archive has grown to include a large number of software packages and modern English corpora, later also the *Helsinki Historical English Corpus*. In persuading Matti Rissanen, the man behind this corpus, to take over the post as Coordinating Secretary, Stig Johansson suggested that the letter 'M' in ICAME should stand for both 'Modern' and 'Medieval' – alternatively, the acronym can now be pronounced with a typically Finnish long consonant: ICAMME.

In this paper I have talked a lot about early 'corpus linguists' and 'corpus linguistics' but not really tried to define what these terms mean. So, in conclusion, here are two quotes which reflect my own view, the first from Graeme Kennedy's useful overview, where you can read about all the other projects I have no time to cover here:

Corpus linguistics is not an end in itself but is one source of evidence for improving descriptions of the structure and use of languages, and for various applications, including the processing of natural language by machine and understanding how to learn or teach a language. (Kennedy 1998: 1)

I also want to express agreement with the view of Wallace Chafe:

What, then, is a 'corpus linguist'? I would like to think that it is a linguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations. Anyone who is not a corpus linguist in this sense is, in my opinion, missing much that is relevant to the linguistic enterprise. (Chafe 1992: 96)

References

- Bolinger, D. (1980), *Language – the loaded weapon, the use & abuse of language today*. London: Longman.
- Carvell, H. T. and J. Svartvik (1969), *Computational experiments in grammatical classification*, Janua linguarum, Series Minor 63. The Hague: Mouton.
- Chafe, W. (1992), 'The importance of corpus linguistics to understanding the nature of language', in: J. Svartvik (ed.) *Directions in corpus linguistics*. Proceedings of Nobel symposium 82, Stockholm 4-8 August 1991. Berlin: Mouton. 79-97.
- Chomsky, N. (1957), *Syntactic structures*. The Hague: Mouton.
- Chomsky, N. (1958), Paper given at *Third Texas conference on problems of linguistic analysis in English*. Austin: University of Texas.
- Crystal, D. (1969), *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.
- Crystal, D. and R. Quirk (1964), *Systems of prosodic and paralinguistic features in English*. The Hague: Mouton.
- Fillmore, C. (1992), "'Corpus linguistics" or "Computer-aided armchair linguistics"', in: J. Svartvik (ed.) *Directions in corpus linguistics*. Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991. Berlin: Mouton. 35-60.
- Francis, W. (1982), 'Problems of assembling and computerizing large corpora,' in: S. Johansson (ed.) *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities. 7-24.
- Francis, W. Nelson (1985), 'After-dinner speech at the 5th ICAME conference at Windermere', *ICAME news*, 10: 5.
- Fries, C. C. (1951), *The structure of English*. London: Longman.
- Greenbaum, S. and R. Quirk (1973), *A student's grammar of the English language*. London: Longman; in the US: *A concise grammar of contemporary English*. New York: Harcourt Brace Jovanovich.
- Halliday, M. A. K. (2005), *Computational and quantitative studies*, volume 6 in *The collected works of M. A. K. Halliday*. Hong Kong: Continuum.
- Jespersen, O. (1938), *En sprogmans levned*. Copenhagen: Gyldendal.
- Kennedy, G. (1998), *An introduction to corpus linguistics*. London: Longman.
- Leech, G. (1991), 'The state of the art in corpus linguistics', in: K. Aijmer and B. Altenberg (eds.) *English corpus linguistics*. London: Longman. 8-29.
- Leech, G. and J. Svartvik (1975), *A communicative grammar of English*. London: Longman.
- Murray, K. M. E. (1977), *Caught in the web of words: James Murray and the Oxford English Dictionary*. New Haven/London: Yale University Press.
- Quirk, R. (1957), 'Relative clauses in educated spoken English', *English studies*, 38: 97-109.

- Quirk, R. (1960), 'Towards a description of English usage', in: *Transactions of the philological society*. 40-61; reprinted as 'The survey of English usage', in: R. Quirk (1968), *Essays on the English language, medieval and modern*. London: Longman. 70-87.
- Quirk, R., A. P. Duckworth, J. L. Rusiecki, J. Svartvik and A. J. T. Colin (1964), 'Studies in the correspondence of prosodic to grammatical features in English', in: *Proceedings of the ninth international congress of linguists*. The Hague: Mouton. 679-691.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1972), *A grammar of contemporary English*. London: Longman.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.
- Quirk, R. and J. Svartvik (1966), *Investigating linguistic acceptability*, *Janua linguarum*, Series Minor 54. The Hague: Mouton.
- Svartvik, J. (1966), *On voice in the English verb*, *Janua linguarum*, Series Practica 63. The Hague: Mouton.
- Svartvik, J. and R. Quirk (1980), *A corpus of English conversation*, *Lund Studies in English* 56. Lund: Lund University Press.
- Svartvik, J. and O. Sager (1978), *Engelsk universitetsgrammatik*. Stockholm: Almqvist & Wiksell.

This page intentionally left blank

Corpus development 25 years on: from super-corpus to cyber-corpus

Antoinette Renouf

formerly the University of Liverpool, UK
(now the University of Central England, Birmingham)

Abstract

By the early 1980s, corpus linguists were still considered maverick and were still pushing at the boundaries of language-processing technology, but a culture was slowly bootstrapping itself into place, as successive research results (e.g. Collins-Cobuild Dictionary) encouraged the sense that empirical data analysis was a sine qua non for linguists, and a terminology of corpus linguistics was emerging that allowed ideas to take form. This paper reviews the evolution of text corpora over the period 1980 to the present day, focussing on three milestones as a means of illustrating changing definitions of 'corpus' as well as some contemporary theoretical and methodological issues. The first milestone is the 20-million-word Birmingham Corpus (1980-1986), the second is the 'dynamic' corpus (1990-2004); the third is the 'Web as corpus' (1998-2004).

1. Introduction

I have been invited to review developments in corpus linguistics over the last 25 years, up to the present day. In fact, given the immensity of the topic, I have decided to focus on corpus creation rather than on corpus exploitation. I shall take the briefest of glances into the future. I am honoured to follow on as speaker from Jan Svartvik, a pioneer in the field in whose corpus-building footsteps I followed only in 1980.

I am no longer the young thing I was when I attended my first ICAME conference in Stockholm in 1982, but I can at least report that I entered corpus linguistics after the first generation of terminology, such as *mechanical recording*, which is such a giveaway to one's age. I was of the generation of UK corpus linguists who by the early 1980s had established that the plural of *corpus* was probably *corpora*, and who, were behind closed doors, decadently using *text* as a mass noun, *data* as a singular noun, and the term *concordances* to mean 'lines within a concordance'.

In this paper, I shall give a brief overview of the history of corpus development, starting with a backward perspective echoing some of Jan's talk, but focussing on the more recent events: the larger 'super-corpora' of the 1980s and 1990s, and the current and future 'cyber-corpora'. I shall refer mainly to the text collections for which I have personally been responsible and am thus better equipped to comment on: the 18 million-word *Birmingham Corpus* (which

evolved into the Bank of English); the open-ended corpora of present-day journalism which my Unit has been processing chronologically over a period of 15 years so far; and most recently, the ‘Web as corpus’, the ad-hoc, always changing corpus of language data extracted from web-based texts. In the course of this review, I shall propose a model to explain the particular path that corpus development has taken, and discuss the theoretical and practical issues involved at different stages of its evolution.

The phases of corpus evolution approximately follow the pattern in Figure 1. The dates there are presented as starting dates since they each refer not just to specific corpora, but to *types*, *styles* and *designs* of corpora which continue to be constructed into the 21st century. The small corpora *LOB* and *Brown* have been followed by their updated equivalents, *FLOB* and *Frown*;¹ the super-sized *British National Corpus (BNC)*² emerged in the wake of the *Bank of English*; the small, specialised *MICASE Corpus of Academic Speech* built on earlier work in spoken corpus production, and so on.

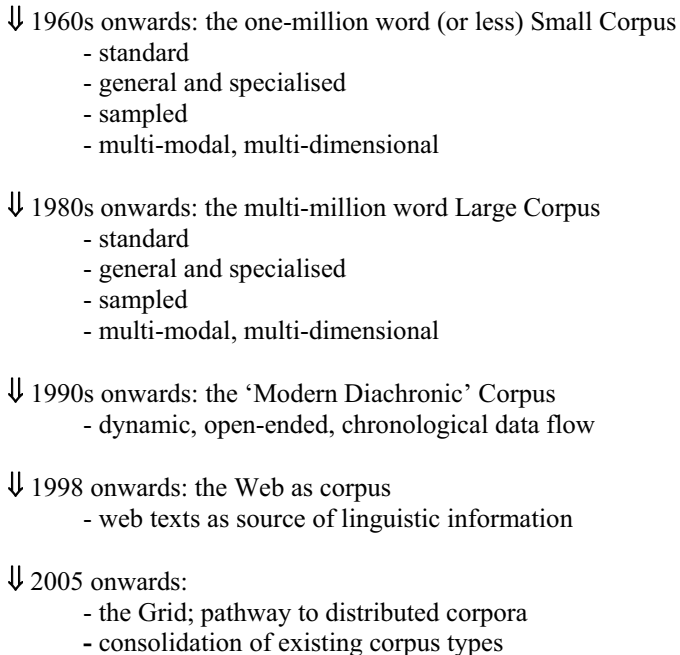


Figure 1: Stages in English language corpus evolution.

2. Major drivers in corpus development

My conception of corpus development is that it has been shaped by three major *drivers*, or motivating forces, over the years. Of course, this is a stylised, simplified representation of what were typically complex and many-layered

circumstances and decisions. These drivers are what I shall characterise for convenience as *science* (or intellectual curiosity), *pragmatics* (or necessity), and *serendipity* (or chance).

The first driver, *science*, is the proper motivation for academics and thinkers. By *science*, I mean the desire to undertake an empirically-based methodological cycle, beginning with curiosity based on introspection, intuition and probably data observation, which leads to the formulation of a research question or hypothesis or conviction, which in turn leads to a programme of hypothesis-testing through data observation, which leads to a discovery, which provides knowledge and experience, which in turn fosters curiosity and the development of further hypotheses to explore.

The second driver in corpus development, *pragmatics*, or necessity, is less exalted. On the one hand, given the time and expense involved in corpus creation, the design decisions taken are inevitably influenced by the availability of particular data, technologies and funds. On the other hand, corpus creators are subject to the vagaries of external agencies, ranging from publishers' requirements, to priorities set by funding bodies, to the exigencies of governmental research grading.

The third driver for corpus development, *serendipity*, is also a powerful one. Jan Svartvik (this volume) reminded us in the context of his creation of the *London-Lund Corpus* that once a concept is demonstrated to be possible, it does not take long for it to be taken up by others. Equally, new data or technology or other innovation can emerge at any moment to support hitherto impossible corpus initiatives. Or vital resources may unforeseeably become available.

Applying the three-category model to the small corpora of the 1960s exemplified above, I would say that the primary driver was *scientific*, and the theoretical underpinnings were as follows:

- that language in use is a bona fide object of study;
- that 1 million words was sufficient to ensure adequacy of grammatical description;
- that exhaustive study was the right approach to a body of text;
- that a corpus could *represent* the language.

More recent small corpora have obviously been built with the benefit of hindsight. My analysis of the continuing creation of small corpora when it is technologically possible to create larger ones is that here necessity is playing a larger role. Many newer small corpora are designed to be comparable with earlier models (e.g. *FLOB* and *Frown*; perhaps learner corpora), some are small because they are very specialised, others are relatively small because of the scarcity of data; e.g. *ZEN* (Fries 1993); the *Corpus of Early English Correspondence* (Nevalainen and Raumolin-Brunberg 1996 and other historical data collections), and many others still are constrained by limited funds and the time pressure to produce results quickly (spoken corpora being particularly costly).

3. The evolution of the super-corpus: 1980s onwards

3.1 Examples of ‘super-corpora’

Birmingham Corpus (1980-1986)

University of Birmingham

20 million words British/American English, written/spoken text

Bank of English (1980-)

University of Birmingham

500 million words British/American English, written/spoken text

British National Corpus (1991-1995)

Longman, Oxford University Press, Chambers

University of Lancaster

Oxford University Computing Services

100 million words British English, written/spoken text

3.2 Drivers for the super-corpus

The same *scientific* curiosity which led to the creation of the small, standard corpora underpinned the development of the first of the next generation of ‘super-corpora’, the *Birmingham Corpus* of 1980-1986. There was continuing curiosity about the nature of real language use, and a desire to discover further unknown facts of the language through as exhaustive study as possible. The difference was that, by the 1980s, there was a realisation that there were questions about lexis and collocation, and indeed even about grammar, which could not be answered within the scope of the small corpus. In addition to scientific curiosity, serendipity assisted the sea change in corpus linguistic thinking. With the emergence of the first corpus-based lexicographic products, perceptions changed within the major publishing houses, and suddenly it became desirable and even indispensable to pay at least lip service to the virtues of corpus-based lexicology. Equally fortuitously, developments in computing technology were creating an ever less hostile climate, providing hitherto undreamed of opportunities for such research.

3.3 Practical issues in the early days of creation of the super-corpus

In the 1980s, whilst the conditions became possible for larger-scale corpus development, there were still formidable practical issues to be overcome. Within the Collins-Cobuild project at the University of Birmingham, Jeremy Clear and I, for instance, had to create the *Birmingham Corpus*, a collection of as near 20 million words as possible, in one year, so that Collins-Cobuild lexicographic work could meet its deadlines. Dashing around like mad things, we were the Ginger (given Jeremy’s colouring) and Fred of corpus building. To begin with, the conversion of printed text to machine-readable form was in its infancy and there were two options for this initial process. The first was to use the so-called Kurzweil Data Entry Machine (KDEM) (Kurzweil 1990), an early optical scanner

the size of a large photocopier, ingeniously designed to convert printed text into Braille. The deciphering capability of this machine was limited in comparison to that of today's scanners, and it tended to offer the character 'W' as a candidate for any character or blob (wryly dubbed the "is it a W?") response, fairly regularly whenever book pages, particularly paperbacks, contained less than perfect print quality. We had two operators working simultaneously non-stop for many months, and to keep up production I had to acquire special dispensation to have women students as well as male working overnight on campus; I processed many books myself. There was a point of diminishing return at which the theoretically slower option, keying-in, became more efficient. This was regularly reached with thinner newspaper pages, through which the KDEM shone a light and diligently tried to read text on both sides simultaneously.

Another practical issue was the acquisition of the texts themselves, and in particular the 'right' edition. The KDEM required two copies of each book, one of which had to be dismembered for scanning, the other required for checking. Books quickly go out of print, and finding two copies of each, dating back to the 1960s, was a headache involving the continual scouring by me of the groves of Birmingham second-handia. The concomitant task of acquiring permission to reproduce the content of hardcopy books in electronic form was a procedure I endured over a 5-year period. The rights manager in most publishing houses was actually a series of individuals who radiated all the joy and urgency of being on fatigue duty. Copyright can be held separately, often by different parties, for the UK, Europe, US and North America, Bermuda, the Cayman Islands, and any combination of these and other remote islands and protectorates. Furthermore, books that are no longer best-sellers and thus present no copyright problems in principle can suddenly be serialised for TV, become overnight successes, and renew their copyright status. In the early days, having to extract these data retrospectively from concordanced output would have caused major logistical problems. Fortunately, only one writer, J. D. Salinger, refused permission for his classic novel *Catcher in the Rye* (1951), unambiguously and in advance, and just a couple of the hundreds of newspaper articles required modest payment for use.

The next issue, text processing, had to be tackled by Jeremy Clear. This was a struggle in the early 1980s. Jeremy progressed from using the punch cards mentioned by Jan Svartvik (this volume) to overnight batch jobs which regularly failed and had to be resubmitted or which, if he was lucky, generated vanloads of output on streams of perforated A3 paper. Jeremy had to contend with the limited processing capacity that was available on early mainframes, of which the first, a UK-manufactured ICL 1906A, was the proverbial walk-in store-room which Jan Svartvik described. According to Jan, Henry Kučera reported that the concordancing of the one-million-word *Brown Corpus* took the total mainframe capacity of Brown University Computer Unit for a day. To process the first-stage 7.3 million-word *Birmingham Corpus*, Jeremy had to commandeer the entire University mainframe resources, and process the data, in 1.2 million word chunks, into 6 batches of concordances, over eight successive weekends. Once he had processed the text, there was still the problem of limited on-screen access to

concordances. In the first year, lexicographers had to work across six sets of microfiches, each alphabetically-ordered. That is, to analyse the word *apple*, they had to look in six places. Of course, there were no PCs, and in-team communications were still paper-bound.

Data storage and processing quickly became easier by 1983, and we were able to move with it, to a larger and more manipulable corpus. Nevertheless, in 1983, the necessity for data compression over larger-scale storage was briefly contemplated.

3.4 Theoretical issues concerning the large ‘general’ corpus

The first issue for the *Birmingham Corpus*, as for the earlier ‘standard’ corpora, was theoretical: how to create a body of text which could be claimed to be an authoritative object of study. Ideally, it would be *representative* of ‘the language as a whole’. But given the unknowability of that totality, and hence the impossibility of designing a perfect microcosm, we justified our design strategies in relation to a network of more readily accessible criteria, such as:

linguistic parameters

- fiction versus non-fiction
- speech versus written text
- authenticity³
- regional and social varieties, domain specificity; generality
- text of particular era and spanning specific time-span statistical parameters
- breadth, variety, ‘principled selection’, relevance and sufficiency
- balance
- sampling, of text extracts versus whole documents; and applying ‘random’ versus ‘non-random’ rules of selection

demographic parameters

- age, gender, social grouping and ethnicity of author
- research needs of corpus users

Though the *Birmingham Corpus* did not exploit them, additional demographic selectional criteria were also available to commercial publishers, such as:

- focus on receiver rather than producer of language
- readership and other sales figures)

Through the 1980s and early 1990s, though it was generally accepted among corpus creators that *representativeness* was unattainable, it was felt necessary to present selectional criteria in those terms. Extracts from some early corpus design rationales are quoted here. Looking back to a chapter in *Looking up* (Sinclair 1987), the companion guide to the *Collins Cobuild Dictionary*, I see that I wrote:

[The *Birmingham Corpus* was] designed to *represent* the English language as it was relevant to the needs of learners, teachers and other users, while also being of value to researchers in contemporary English language. (Renouf 1987: 2)

while Della Summers (1993) said of the *Longman/Lancaster English Language Corpus* that it was:

representative of the standard language in a very general sense, not restricted to a regional variety (e.g. British English or a local dialect) or a narrow range of text types (e.g. scientific texts, newspaper writing, language of a particular social class). (Summers 1993: 186)

Concerning the *BNC*, text selection was characterised in the following terms:

text that is published in the form of books, magazines, etc., is not *representative* of the totality of written language that is produced. [...] However, it is much more representative of written language that is received, and... thus forms the greater part of the written component of the corpus. (Burnard 2000: 1)

Meanwhile, Matti Rissanen (1992) acknowledged, of the *Helsinki Corpus of English* that:

Just as a corpus will never reliably reflect the language in all its varieties and modes of existence, so, too, parameter coding can never hope to give a complete and theoretically valid description of the samples. (Rissanen 1992: 188)

As the *BNC* was emerging, the achievability of true *representativeness* was historically debated by John Sinclair and Willem Meijs at a 1991 conference in St. Catherine's College, Oxford. This grail continues to haunt corpus linguists as we progress into the new millennium. See, for example, even as recently as 2004, the stated ambitions for the new *Corpus of Spoken Israeli Hebrew* (CoSIH) at Tel Aviv University (Izre'el and Rahav 2004) are as follows:

Our aim is to produce a *representative* sample which will take into account not only demographic criteria but also account for contextual varieties. Thus, data should be collected according to two distinct types of criteria: while sampling is conducted according to statistical measures and thus will be *quantitatively representative* of the entire population, collecting data according to analytical criteria is not necessarily *statistically representative*; the *representativeness* of a corpus must be based on the internal social structure of the speech community for which it is designed. (Izre'el and Rahav 2004: 7)

Stepping back, the only progress that I have noticed in this debate over the last twenty-five years, which is presented each time as if it were a new problem, is a change in the terminology used to conduct it! According to our *WebCorp* output, the variant form *representativity* is now in vogue:

1. There is now a debate among corpora research community concerning this concept of the **representativity** of the data gathered inside text corpora
2. So it seems that we now witness a move from **representativity** towards reusability
3. How can we measure the **representativity** of a corpus with respect to a given linguistic construct?
4. That raises the problem of the **representativity** of the data base and of the application of methods for the presentation of findings.
5. We consider issues such as **representativity** and sampling (urban-rural, dialects, gender, social class and activities
6. In search of **representativity** in specialised corpora
7. The discussion of issues of corpus annotation, the **representativity** of corpora, economy, and an optimized structuring of the data
8. Their **representativity** is measured by reference to external selection criteria
9. what is meant by 'spoken' and 'written', the **representativity** of a certain type of language
10. The twin theoretical problems of data **representativity** and corpus structuration.
11. the difficulty of defining **representativity** in such a corpus.

Figure 2: *WebCorp* output for the term *representativity*, May 2004.

A fundamental tenet of corpus linguistics in the 1980s (Johansson 1982) was breached by the larger corpus, which was that a corpus should be studied exhaustively. The purpose had been to ensure that the linguist exploited a small resource exhaustively, so as not to miss any aspects of language use of which he/she had not been aware, and with a view to applying quantitative measures to establish the relative significance of phenomena in the corpus. The small corpus was a precious, hard-won object. The advent of larger corpora necessitated new analytical approaches. While some phenomena remained sparse even in the huge text collections, many more words were well represented, and some of the commonest phenomena in text, the grammatical and phraseologically productive lexical items, now occurred in such vast numbers that they could not always be studied exhaustively, but had to be sampled (as in the Cobuild lexicographic project, for instance).

3.5 Was building the large *Birmingham Corpus* worth the effort?

Yes, it was. It revealed the nature of lexis and collocation, and underpinned the ground-breaking *Collins-Cobuild Dictionary* as well as countless other books and theses then and since. Importantly for subsequent applied research, it also

revealed the relationship between surface patterns of text and meaning. The paradigmatic axis of language only being realisable syntagmatically in text, the link between collocation and word meaning is readily identifiable at the surface level. In the course of time, the awareness of this fact allowed me to devise and set up some large-scale projects – AVIATOR⁴ (Birmingham 1990-1993, ACRONYM⁵ and APRIL⁶ (Liverpool 1994-2000) – in which, in collaboration with a series of inventive software engineers, including Alex Collier, Mike Pacey and now Andrew Kehoe and Jay Banerjee, I have been able to exploit the regularity of word collocation in text to the limits. In AVIATOR, for instance, the change in meaning of a word was identified automatically by a significant change in its collocational patterning in a corpus of journalism. In ACRONYM, by dint of the simple comparison of the collocational profiles of words, we can find synonyms and near synonyms ('nyms') automatically. In 2004, ACRONYM produced the results, shown in Figure 3 for the word *solace*.

comfort	excuse
consolation	happiness
inspiration	cure
satisfaction	counselling
encouragement	respite
reassurance	salvation
warmth	succour
refuge	sanctuary
shelter	remedy
punters ⁷	haven

Figure 3: ACRONYM extract of ranked 'nymic' output for the term *solace*, *Independent* news text, 2004.

Moreover, in 1996 for the word *surgery*, it produced the multi-word 'nyms' shown in Figure 4.

heart transplant	coronary event
heart surgery	median survival
heart transplantation	surgical procedure
reduction surgery	surgical resection
breast biopsy	surgical intervention
coronary angioplasty	artery bypass
coronary angiography	outpatient surgery
coronary revascularization	prophylactic mastectomy
coronary stenting	angioplasty procedures

Figure 4: ACRONYM extract of ranked multi-word 'nymic' output for the term *surgery*, *Independent* news text, 1996.

These were clearly promising first-stage results, unedited as they are, in themselves raising important questions about the nature of lexical semantics in use, and having enormous potential for applications in fields from lexicography to database management.

4. The evolution of the modern diachronic corpus

The concept of a diachronic, or ‘monitor’, corpus had been raised as a theoretical possibility back in 1982 by Sinclair (Johansson 1982). His vision had been of language as a changing window of text, in which the computer would ‘monitor’ aspects of language use across time, but would then discard each stretch of text once processed.

It was not until 1990 that the first ‘dynamic’ corpus of unbroken chronological text was finally established by the RDUELS Unit in our 1990-1993 AVIATOR project at Birmingham, using text from the *Times* newspaper dating back to 1988. By that time, there was no obstacle to its being archived in its entirety, thus allowing the flexibility to return to the data as new hypotheses emerged for examination. The second ‘dynamic’ corpus of this kind was set up by the RDUES Unit in the ACRONYM project at Liverpool in 1994, this time with *Independent* news text, also starting from 1988, the date of the inception of that newspaper.

There are three types of corpus which currently support diachronic study to the present day, each representing different approaches to diachronic study. The first type comprises the small, synchronic but parallel ‘standard’ corpora, *Brown*, *Frown*, *LOB* and *FLOB* (Brown University, Universities of Lancaster, Oslo, Bergen and Freiburg); the second type is the chronologically-ordered corpus of text samples of historical English register now reaching into the 20th century, known as the *Archer Corpus* (Universities of Arizona, Southern California, Uppsala and Freiburg). The third type is represented by the aforementioned unbroken, chronological data flow of *Times*, and more recently of *Independent* and *Guardian* journalistic text (now at the University of Central England, Birmingham).

4.1 Drivers for the modern diachronic corpus

The motivation for the development of the modern diachronic corpus was primarily *scientific*, or theoretical. It manifested itself in the following ways:

- awareness that language is a changing phenomenon;
- belief that language change can in principle be observed in corpus data;
- curiosity about innovation, variation and change in grammar and lexis.

In our case, it involved a curiosity about neologistic assimilation, lexical productivity and creativity; and the structure of the lexicon at the level of hapax legomenon. I was also curious to investigate the power of collocation to identify

change in word use, in word sense and in meaning relationships between words. Supporting this theoretical impetus for the modern diachronic corpus was the serendipitous availability of news text held as an electronic flow, and of the necessary funding.

4.2 Theoretical issues concerning the ‘dynamic’ diachronic corpus

There are fundamental differences between static and dynamic corpora, both in purpose, design and in methodology. The purpose of a dynamic corpus is to support the study of language change over time. Monitoring chronologically-held text reveals innovation, trends and vogues, revivals, patterns of productivity and creativity. In return, however, the goal of quantification and thence the assignment of significance which was taken for granted with the static, finite corpus becomes impracticable with the open-ended flow of text, since it requires knowledge of the size of the total population.

Different theoretical approaches are taken to the study of modern diachronic corpus data. The clearest distinction lies on the one hand between the study of language change across a significant period in two parallel finite corpora, such as *Brown* and *Frown*; and on the other, the ongoing study of change in an open-ended, unbroken flow of data across 10-15 years, as with my Unit’s work.

The parallel finite corpora which are currently available sit thirty years apart, a span which has been argued (Mair 1997) to be appropriate to reveal significant shifts in language use. Under these conditions, Mair and others (see Hundt 2006; Leech and Smith 2006) have, in addition to other insights gained, for example into American and British English variation, been able to investigate and substantiate their theories of language change in relation to grammaticalisation, importantly linking this with increased colloquialisation across the period. A limitation of the ‘gapped’ approach, as pointed out by Mair, is the improbability of capturing actual moments of change (at least, conclusively). Conversely, the unbroken diachronic corpus of news text has the disadvantage that it covers a time-span too brief to reveal much if anything of grammatical change, but it does reveal lexical and morphological innovation and fashion, and hints of change at the lexico-grammatical levels. It is more likely to capture many actual moments of journalistic invention (though again unverifiably), as well as the early quotation and exploitation of neologisms.

The notions of studying change and of processing text chronologically, tracing neologistic activity, were and probably still are considered significant philosophical and methodological steps forward. However, by the 1990s, the computational capacity to store and retain all past data led to a new theoretical breakthrough and associated methodologies. It allowed the bi-directional processing of text,⁸ and thus the analysis of character strings as multiply segmentable. This breakthrough was explored in the APRIL project, 1994-1997. The project concerned the identification and classification of hapax legomena and other neologisms at point of first-occurrence (all singletons at that stage) in the journalistic text flow, with the purpose of understanding the nature of the lexicon at its most productive level. It became clear that while a new word sometimes

consisted of existing words or affixes in new combinations, the parse for new formations yet to appear in dictionaries very often revealed ambiguity. Multiple analyses of the character string for two or more possible points of segmentation were necessary. Words could not simply be parsed from left to right, but required crab-wise incremental assessment, whereby the word was regarded as potentially constituted of a series of partially overlaid sub-strings, and these analytical possibilities were ranked by ‘cost’ or likelihood. Such recursion was now achievable with the application of a novel ‘chart parser’, modified to operate at character level.

A host of terminological issues accompanied the move into diachronic corpus study. An early question concerned what the terms *corpus*, *database* and *text archive*. By 1990, a *corpus* was still a designed, finite collection of partial or whole raw texts (if annotated, then an *annotated corpus*), and processed electronically, typically as a synchronic entity. A *database* could encompass a *corpus*, but typically denoted a collection of knowledge or facts, the products of raw corpus analysis; and a *text archive* was a catalogued text collection, whether or not prepared for processing as a corpus. Inevitably, there remains some overlap in the use of these terms.

Another terminological distinction exists between *historical* and *diachronic* (Renouf 2002). Historical linguists use the term *diachronic* to describe their study in a collective sense, since as a body of scholars, they study the whole realm of text across time, even though in principle each individual investigation could be focussed synchronically in the past. To modern diachronic linguists, *diachronic* is used in contrast to *synchronic*; while *diachronic* can also mean ‘across time’ to both historical and modern corpus linguists.

Moreover, when historical linguists speak of *diachrony*, their mental time-frame is likely to be calibrated in centuries and millennia. Modern diachronic linguists, in contrast, are typically comparing text across time-frames of ten to thirty years. Rissanen (2000) has referred to the longer time-frame as ‘long diachrony’; Kytö, Rudanko and Smitterberg (2000) have talked of the shorter time-frame in terms of ‘short-term change in diachrony’; while Mair (1997) has dubbed this shorter time-span ‘brachychrony’.

The gradual coming together and overlapping of the periods of text studied by historical and modern corpus linguists has also given rise to a terminological lacuna for the period extending from 1900 to today. ‘Late Modern English’ would normally seem an appropriate term, but of course this means something much earlier to the historical linguist. In fact, corpus linguists refer to this nameless period as anything from ‘modern’ to ‘present-day’ to ‘current’ to ‘20th century’ and ‘21st century’.

4.3 Practical issues concerning the modern diachronic corpus

For some types of modern diachronic corpus study, such as that based on the comparison of two parallel designed (‘standard’) corpora like *Brown* and *Frown*, the parameters of corpus text selection can be set according to specific linguistic criteria. For diachronic study within an unbroken flow of text, the nature of the

corpus is dictated by the electronic data which are available; in the AVIATOR, ACRONYM and APRIL projects, this was quite simply national news text, although this did not prevent us, for example, from branching out to compare English and French news streams. The scope of the diachronic corpus is also dictated by the time-span of available data. In 1990, we in my Unit began the AVIATOR project with a store of just 5-6 years of text; now in 2004, we have over 15 years' worth of *Independent* and *Guardian* news text. Another problem is that the phenomena that are typically under scrutiny in modern diachronic study, such as first occurrences, neologistic usages and new patterns, dribble through in ones and twos when analysed diachronically. Very delicate statistical measures may be introduced gradually, but their performance on sparse early data has to be scrutinised closely.

4.4 Was building the modern diachronic corpus worth the effort?

Yes, it was. The monitoring of text reveals innovation, trends, vogues, changes, revivals, patterns of productivity and creativity. In the ACRONYM project, for instance, we have been able to discover changes in sense relations in text over time by monitoring the change of collocational profiles of two words. For the word *cleansing*, we find an upsurge of activity accompanied by a change in sense relations (and thus incidentally reference) in the latter part of 1992 in *Independent* news text. This is shown in Figure 5.

ind8901-ind9203

inquiry
centre

ind9204-ind9912

massacres	slaughter
genocide	war
atrocities	offensive
killings	refugees
expulsions	shelling
repression	bombardment

Figure 5: Ranked list of 'nyms' for the term *cleansing* in *Independent* news text, July-December 1992.

Meanwhile, in the APRIL project, the analysis of diachronic data allows us for the first time to trace over time the neologisms formed with particular productive affixes. Figure 6 illustrates graphically a significant growth in the frequency of neologisms formed with the prefix *e-* over an eight-year period.

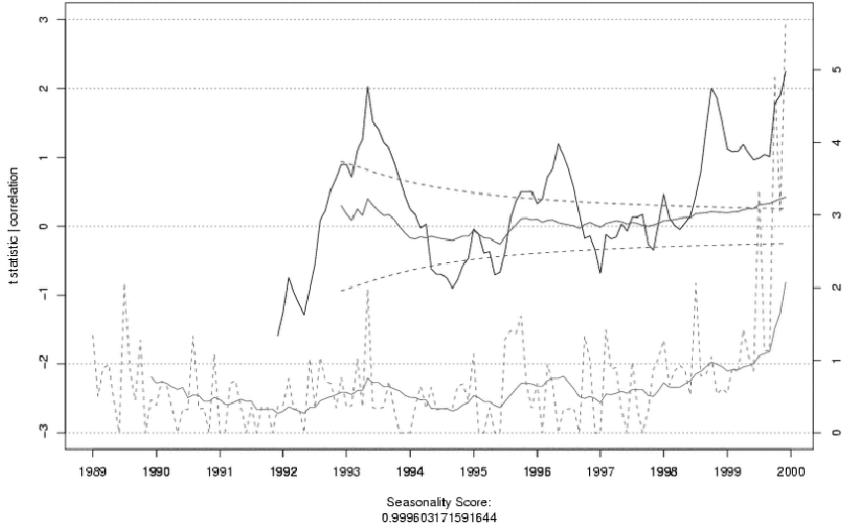


Figure 6: Graph showing significant growth in the number of new words with the prefix 'e-' in *Independent* news text, 1992-2000.

In Figure 7, we illustrate this with some output for the prefix *cyber-*. The data are the result of morphological analysis by a specially-modified 'chart parser'. By focussing at sub-word level in our diachronic text, we have a means of grouping and classifying these hapax legomena of text. In Figure 8, we demonstrate how it is then possible to extrapolate from the same diachronic corpus of news text an indication of the ranking and relative growth (for the prefixes *techno-* and *cyber-*) and drop (for the prefix *poly-*) in productivity for affixes over a period of ten years. These kinds of cumulative analysis provide us, in principle, with a basis for the prediction of the overall structure of the lexicon.

Query Results from *The Independent* - All of 1999

New words with the prefix 'cyber'

word	parse	tag	month
☐ - cyberstalker	cyber- (stalker)	NN1	9901
☐ - cyber-future	cyber- '-' (future)	NN1	9901
☐ - cybermemoir	cyber- (memoir)	NN1	9901
☐ - Cybersouls	cyber- (souls)	NN2	9901
☐ - Cybertalk	cyber- (talk)	NP1	9901
☐ - cyber-event	cyber- '-' (event)	NN1	9901
☐ - cybergeek	cyber- (geek)	NN1	9901
☐ - cybermemoirists	cyber- (memoirists)	VVZ	9901
☐ - cybersong	cyber- (song)	VVG	9901
☐ - cyber-sessions	cyber- '-' (sessions)	NNT2	9902
☐ - cyber-glow	cyber- '-' (glow)	NN1	9902
☐ - cyber-realist	cyber- '-' (realist)	NN1	9902
☐ - Cyber-Valentine	cyber- '-' (valentine)	NP1	9902
☐ - cyber-commerce	cyber- '-' (commerce)	NN1	9902
☐ - cyber-auction	cyber- '-' (auction)	NN1	9902

Figure 7: Neologisms with the prefix 'cyber-' in *Independent* news text, 1999.

	89	90	91	92	93	94	95	96	97	98	99
non	1	1	1	1	1	1	1	1	1	1	1
un	1	1	1	1	1	1	1	1	1	1	1
anti	1	1	1	1	1	1	1	1	1	1	2
ex	1	1	1	1	1	1	1	1	1	1	2
re	1	1	1	1	1	1	1	1	1	1	2
pre	1	1	1	1	2	1	2	2	2	2	2
post	1	2	2	2	1	1	2	2	1	2	2
over	1	1	1	1	2	2	2	2	2	2	2
euro	1	2	2	2	2	2	2	2	2	2	2
super	2	2	2	2	2	1	2	2	2	2	2
mini	2	2	2	2	2	2	2	2	2	2	2
pro	2	2	2	2	2	3	2	2	3	3	2
semi	2	2	2	2	3	3	3	3	3	3	3
out	2	2	3	3	3	3	3	3	3	3	3
sub	2	2	3	3	3	3	3	3	3	3	3
under	2	2	3	3	3	3	3	3	3	3	4
multi	3	3	3	3	3	3	3	3	3	3	4
mid	2	3	3	3	4	4	3	3	3	4	4
pseudo	3	3	3	4	4	4	3	3	4	4	4
quasi	3	3	3	4	3	3	4	4	4	5	4
mega	4	4	4	4	3	3	3	4	4	4	4
inter	2	3	3	4	4	4	4	4	4	5	5
micro	3	4	4	4	4	5	4	3	4	4	4
ultra	4	4	4	4	3	4	4	4	4	4	4
counter	3	3	4	4	4	5	5	5	5	5	5
eco	4	5	4	4	5	5	5	4	4	4	5
hyper	3	5	5	5	5	5	4	4	4	4	5
mis	2	3	4	4	5	5	5	5	5	5	6
neo	3	4	4	4	5	5	5	5	5	4	5
auto	4	4	5	5	5	4	5	5	4	5	5
bio	4	5	4	5	5	5	5	5	5	4	5
mock	4	4	5	5	5	5	5	4	5	5	5
dis	3	4	5	4	5	5	5	5	6	6	6
techno	6	7	6	6	5	5	3	3	4	4	5
tele	5	5	5	5	5	5	5	5	5	5	6
cyber	8	9	8	9	8	5	3	2	2	2	3
arch	5	5	6	6	5	7	5	6	6	5	5
trans	4	6	5	6	5	5	6	5	5	7	7
proto	5	5	6	7	6	7	5	5	6	5	5
poly	4	5	5	5	6	7	6	6	6	6	7

Figure 8: Prefix banding of productivity for the prefixes *techno-*, *cyber-* and *poly-* in *Independent* news text, 1989-1999.

5. The evolution of the cyber-corpus

I use the term ‘cyber-corpus’ to refer specifically to texts on the World Wide Web treated as an on-line corpus, in the sense of functioning as a source of language use, of instances of language use extracted from the Web and processed to provide data similar to the concordanced and other analysed output from a conventional corpus.

5.1 Drivers for the Web as corpus

There are three drivers for the treating the texts on the Web as a source of linguistic information. The primary one is probably *serendipity*. The Web emerged in the 1990s, and though it was created to store textual and other information, people gradually realised that it contained useful and otherwise unavailable linguistic data, new and rare words not found in existing corpora, as well as more varied types of language use. *Pragmatic* considerations were the secondary driver. Corpora are very expensive and time-consuming to build, so that they are limited in size, out of date by the time of completion, and they do not change or keep up with language change. Web texts, on the other hand, are freely available, vast in number and volume, constantly updated and full of the latest language use. *Theoretical* curiosity about the nature and status of rare, new, and possibly obsolete language phenomena had been nurtured by many a linguist prior to the birth of the Web, and as web texts began to accumulate, corpus linguists began to negotiate the commercial search engines in an attempt to search for instances and counter-instances to test their new or long-nurtured hypotheses. So the demand was there, and as with the opening of any new highway, demand increased simply in response to its existence.

5.2 Theoretical issues concerning the Web as corpus

The theoretical objections to using the Web as a corpus come thick and fast. The problem is exacerbated where the processing, as with *WebCorp*, is carried out in real time. One of the many issues is the uncontrollability of the data, which form an arbitrary and instantial corpus that changes like the sand with each new search. Another is that fundamental corpus-linguistic methods such as exhaustive and quantitative study are impossible or inhibited in this non-finite context, where the total population is not known or knowable, and the significance and interpretability of results are thrown into question. A fuller range of problems has been explored, e.g. in Kehoe and Renouf (2002), Renouf (2003), Renouf et al. (2006), and is echoed in the practical issues enumerated below.

5.3 Practical issues concerning the Web as corpus

Many linguists have not yet reconciled themselves to the advantages of accessing the Web as a corpus, finding all manner of objections to it in principle. Chief among their concerns are probably the heterogeneity and arbitrariness of the text

and hence the status of the language use found on the Web. These concerns are shared by us as active web-as-corpus providers, who struggle to extract from the Web and process them to a degree which at least approaches the interpretability and usability of the output from conventional diachronic corpora, but which currently fails to match the quantifiability of conventional finite corpora. Researchers like ourselves cope with such issues as the state of Web text, with its typographical errors and erratic or absent punctuation (useful for sentence identification); the heterogeneity of web-held data, the handling of web pages with their hotchpotch of more and less text-like texts; the need for speed of search, retrieval and processing; language identification; and the absence of an established standard for dating texts, with particular reference to the date of authorship (whence also the impossibility of achieving reliable chronological text sequencing for diachronic study).

5.4 Was building the Web as corpus worth the effort?

Yes, it was. *WebCorp* gives us the possibility of retrieving instances of words and phrases in text that are either too rare or too recent to appear in conventional text corpora. Figure 9 illustrates the case with reference to the term *Enron*, which appeared on the Web as soon as the scandal broke, and almost immediately became productive, spawning not just the search term *Enronomics*, but also *Enronyms*, *Enronitis*, *Enronity*, *Enronethics*, *Enronizing*, *enronish*, *Enronitize* and *enronomy*, to name just those variants in our sample.

WebCorp output for search term "Enronomics"

Domain: ".uk OR .com"

1. attack Bush's economic policies with the term "[Enronomics](#)" (a phrase that apparently originated in a
2. to Believe He Knows About the Economy? [Enronomics](#) = Contributors Get Richer 1/16 Message to
3. corporate malfeasance. Recently spotted Enronyms: Enronitis, Enronify. [Enronomics](#). silver bullet: In war, it's an
4. is laid bare by what rivals call '[Enronomics](#)' - the political fable of the Enron corporation
5. Dems slogan for slogan and neutralize the [Enronomics](#) accusations, may I coin the term "Enronethics
6. C.) 2 p.m. Breakout Workshops -- Confronting [Enronomics](#) -- Arianna Huffington, Rep. George Miller (D-Calif.)
7. investigators Wed. (DBN Subscription Required) Democrats knock '[Enronomics](#)' - But strategists warn against
8. a political problem for TeamBush-with talk of "[Enronomics](#)," or "Enronizing" Social Security and Medicare. But
9. believing their press, watch out. It's [Enronomics](#), folks. The rich seducing the poor, while
10. Riseth The Conservative Cliterati What Monica Cost [Enronomics](#) Catholics and Condoms Virtual Rape Ring-a-Ding
11. national energy policy based on the same [Enronomics](#) as its own disastrous business strategy. But
12. people, to be enronish and to practice [Enronomics](#). "We've seen ugly, enronish sights before," Jane
13. The Looting of America: Reaganomics, Clintonomics and [Enronomics](#) AL MARTIN is America's foremost
14. strategy?" http://www.dailyhowler.com/th012902_1.shtml [Enronomics](#) Explained (deliberately driving the country into
15. who's spent two weeks talking about Bush's "[Enronomics](#)" and "Enronizing" Social Security. He capitulated to
16. McFedries said. He placed worse odds on "[Enronomics](#)," reminiscent of "Reaganomics," sticking. "The Democrats
17. ideology. It blows the lid off Bush's [Enronomics](#), and his plan to Enronitize Social Security
18. at alarming rate. ENRON (I call it [Enronomics](#)) phenomena - soft money manipulating the policies and
19. can only be described as trickle down [enronomics](#)! That's exactly what Bush is doing to
20. hardest hit by the Bush trickle down [enronomics](#). Now it looks like the Bush enronomy

Figure 9: *WebCorp* output showing the productivity of *Enronomics*, May 2004.

WebCorp output for search term “medicalisation”

1. legislation was shifted from criminalisation to medicalisation of drug use. Public demand for effective treatment
2. Sheldon examines the causes and effects of the medicalisation of abortion, focusing on the role that law
3. to the lawfulness of the procedure remains. The Medicalisation of Abortion and the Common Law The
4. celebrated 6. 02 6620 2970 Fax: 02 6620 2161 The Medicalisation of Sexual Violence: The Social and Political
5. their frustration and disappointment with the increasing medicalisation and intervention in maternity care and
6. began which medicalise, and therefore pathologise, difference. The medicalisation of epilepsy has inevitable
7. 28. Grubb, A,;" Abortion Law in England: The Medicalisation of a Crime", 18 Law, Medicine and Health
8. School: School individuals and society; autonomy and paternalism; the 'medicalisation' of life; the goals of
9. medicine; the society is: adapt yourself" (Touraine) The psychologisation / medicalisation of school education
10. is a strategy for to turn back the tide on the 'medicalisation' of everyday life. People, who would previously
11. upon initially vague ideas of clinical stress, medicalisation of tension, emotional fever, & on perceptions about
12. and death and to exert control, in line with most medicalisation of childbirth, but Petchesky points out that
13. violation and we refuse to see the medicalisation of something that is wholly unnecessary. We Sociology,
14. Brandeis University, Massachusetts Medical sociology The medicalisation of deviance Modern genetics Email:
15. Crawford, E on midwifery knowledge before the NHS, the medicalisation of childbirth and on the teaching of
16. not to be underestimated. [34] A danger of medicalisation of the law becomes apparent, for instance, in
17. wards etc Centre, Oxford University. Leading critique of the medicalisation of distress via the diagnosis of
18. PTSD the postcolonial condition citizenship and rights discourse medicalisation of the legal subject revisiting
19. consent AIDS A political sociology of lifestyle pharmaceuticals and medicalisation. University of Sussex Mr
20. MM Hopkins THE Midwifery. Professional education Fiona Dykes Infant feeding. Medicalisation of childbirth
21. Norma Fryer Ethical issues relating of mental illness. Experts call it the "medicalisation of human distress" –
22. the trend to treat professionalisation' of lay researchers and representatives. The Medicalisation of Health and
23. Use of lethal injection and the general medicalisation of killing are in direct conflict with Field D (1994)
24. Palliative medicine and the medicalisation of death, European Journal of Cancer Care Recentering Class and C

Figure 10: *WebCorp* output showing old and new uses of *medicalisation*, 2004.

Figure 10 also exemplifies the richness of web text in terms of the changing use of words across time which it (inadvertently) yields. Here, the term *medicalisation*, in the new sense of ‘treating as a medical condition the natural facets of life’, is found alongside the established meanings of ‘decriminalising (of drug use)’ and ‘treating terminal illness by palliative means’. In addition, this text shows how a derived term like *medicalisation* can spawn parallelisms, here *pathologise*, *psychologisation* and *professionalisation* (or vice versa). Searching the Web with the aid of wildcards and brackets, combined to represent complex patterns, yields another dimension of valuable information. Figure 11 presents an extract of the pattern e.g. *dr[i]o]ve[s|n] * [a]round the*.

1. Start up *drives me round the twist*
2. Fury over lorry that *drives residents round the bend*
3. Over used, that stupid drumbeat *drove me round the bend*
4. We quit – you’ve *driven us round the bend*
5. The noise *drove her around the bend*
6. Her Majesty was *driven twice round the Mews yard*
7. ‘Sick’ Diana pic *drives critics round the Benz*

Figure 11: *WebCorp* results for pattern *dr[i]o]ve[s|n] * [a]round the*, using wildcard option and bracketing.

6. A possible future of the corpus in corpus linguistics

While the range of small, medium and large corpora will undoubtedly continue to grow, it seems likely that the World Wide Web will also figure largely in future corpus development, both as a source of data (whether online or downloaded) in itself, and also as a repository for datasets and text collections, a staging post in the location and access of externally-held intranet text archives and corpora, and a medium in which researchers can cooperatively create, process and share corpora. This prediction is based on discussion that has been ongoing for several years now concerning the post-Internet era of 'GRID'-like technology.⁹ The GRID¹⁰ refers to a set of new hardware that will support, in principle cooperatively, a more distributed processing environment.¹¹ To the corpus linguist, the new web infrastructure, as it gradually appears, may not seem much different from now.¹² The new structure will probably be a layer of protocols or routines sitting on top of existing web architecture which function in a similar way to already familiar protocols such as 'ssh' (secure shell, for using a computer remotely); 'email' (Pine, Mulberry etc, for sending electronic messages); and 'ftp' (file transfer protocol, for copying files over networks).

6.1 Drivers for the corpus of the future

As has been implied, the driver for this development towards co-operatively processed corpora will be a *pragmatic* one. As the scale and type of computing increases, there will be a growing need for regionally distributed computing, for information and resource sharing, and of handling the logistics of moving around distributed data and other resources.

7. Concluding remarks

I have in this paper traced the development of corpora 'twenty-five years on', reviewing the issues involved at each stage. I have shown that as the years have passed, the design of corpora has continued to be characterised by the tension between the desire for knowledge and the constraints of practical necessity and technological feasibility. The vision of corpus evolution that I have presented, being a chronological overview, has tended to coincide with major milestones in technological evolution. This is inevitably a simplified picture, which has not given due attention to other factors. Corpus linguists themselves, while still curious about real language in use, have been growing increasingly aware of what it is, and aware that every type of corpus is in principle now available to them, or possible to construct.

As computing technology hurtles on, it does open up access to new corpus resources and the possibility of new orders of corpus magnitude such as I have inventoried, but it also provides the infrastructure to support smaller-scale, more intricate corpus design and organisation, exemplified by the multiply-layered, cross-disciplinary databases such as the *Corpus of Early English Correspondence*

(Nevalainen 2003), the *Corpus of Early English Medical Writing* (Taavitsainen 1997) and the *LAOS (Linguistic Atlas of Older Scots)*. And much research activity has not evolved directly in response to the availability of large-scale computing resources but to the linguistic needs of the corpus linguist. The consolidation of existing corpora is, for example, the focus of much attention. Synchronic corpora such as the *BNC* are being renewed or updated; diachronic and variational collections such as the *Archer Corpus* are growing incrementally; corpus sets like the *ICLES* learner corpora and the *LOB, FLOB, Brown, Frown* stable are being extended.

Ultimately, each of these corpus-oriented activities, whether ground-breaking or incremental, computational or linguistic, theoretical or intuitive, progressively enriches our understanding, whilst broadening the scope of enquiry for the next phase.

Notes

- 1 Thus, when I speak of the early 'small corpora', I am referring primarily to the pioneering work of the first generation of corpus linguists in the form of the *Brown Corpus* (1961-1967), *Survey of English Usage* (1953-1987) *Corpus* (1964-), *Lancaster-Oslo-Bergen Corpus* (1960-1967) and *London-Lund Corpus* (1970-1978); and not so much to the important but second-wave small corpora of specialised text, which were created from 1980 onwards, notably those of historical English such as the *Helsinki* and *Archer* corpora; the corpora of regional varieties of English, including Somerset/Norfolk and Manx English dialects; the corpora of international varieties of English such as those of the *ICE* Project; the corpora of Learner English such as those of the *ICLE* and *CLEC* projects; the multilingual corpora such as those developed in Sweden and Norway; or the specialised technical corpora, such as *MICASE Corpus*.
- 2 The *BNC* was created in 1991-1994 by an academic-industrial consortium whose original members were: Oxford University Press, Longman Group Ltd, Chambers Harrap, Oxford University Computing Services, the Unit for Computer Research on the English Language (Lancaster University), British Library Research and Development Department.
- 3 Authenticity was considered essential, though initially what precisely constituted authentic text, and what its benefits were to language description, were not initially questioned in detail. Sinclair remedied this, in his article on 'naturalness' in language, published along with a reply by Owen, in a useful volume on the topic (1988). Meanwhile, whilst it was clear to most corpus creators from the outset that dramatic dialogue did not constitute an authentic use of speech, and was thus to be excluded, the issue of how to classify conversation within novels has remained largely unresolved.

- 4 AVIATOR stands for ‘The Analysis of Verbal Interaction and Text Retrieval’.
- 5 ACRONYM stands for ‘The Automated Collocational Retrieval of Nyms’.
- 6 APRIL stands for ‘The Analysis and Prediction of Innovation in Text’.
- 7 The semantically unrelated word *punters* creeps into the nymic output by virtue of the unexpected number of collocates it shares with *solace*. It is not possible by statistical means to avoid such oddities entirely, only to find the best variables for minimising their occurrence.
- 8 The use of short word histories, in the form of tri-grams where two recognised words to its left were used to recognise a third unknown node word, preceded this work (Jelinek 1976), but this was in the era of off-line processing of static finite corpora, rather than the new ‘on-the-fly’ processing of our approach.
- 9 Particle Physicists devised the World Wide Web, and facing many of the problems the ‘Grid’ aims to solve, took the first initiative.
- 10 Other projects which resemble the Grid, such as Internet II and Internet III, exist internationally.
- 11 The term GRID is “chosen by analogy with the electric power grid, which provides pervasive access to power”, a facility which the user simply plugs into, without knowing where the electricity is processed or comes from, and which “has had a dramatic impact on human capabilities and society” (Foster and Kesselman 1999: xix). If and when the Internet saturates, it will need to be re-organised. The result will probably be that the Web contains more indexes to texts than texts themselves, which will be stored in ‘DataGrid intranets’. Specialised software will be needed to process them within a predicted system of global distribution of machinery (from regional computing hubs down to domestic machines), and ‘Grid middleware’ will be required to find and organise them, and to protect security. This new system will also require better text content annotation, of the kind being devised by the ‘Semantic Web’ community of computational linguists.
- 12 The definition of large-scale processing to the particle physicist envisages something on a vastly larger scale than that of even the most ambitious corpus linguist.

References

- Burnard, L. (ed.) (2000), *Reference guide for the British National Corpus (world edition) (BNC CD-ROM)*. Oxford: Humanities Computing Unit of Oxford University.

- Foster, I. and C. Kesselman (eds.) (1999), *The Grid: blueprint for a new computing infrastructure*. San Francisco: Morgan-Kaufmann.
- Fries, U. (1993), 'ZEN-Zurich English newspaper corpus', in: M. Kytö, M. Rissanen and S. Wright (eds.) *Corpora across the centuries. Proceedings of the first international colloquium on English diachronic corpora*. St Catharine's College, Cambridge. Amsterdam/Atlanta: Rodopi. 17-18.
- Hundt, M. (2006), "'Curtains like these are selling right in the city of Chicago for \$1.50": the mediopassive in 20th-century advertising language', in: A. Renouf and A. Kehoe (eds.) *The changing face of corpus linguistics: papers from the 24th international ICAME conference*, Guernsey, 23-27 April 2003. Amsterdam/New York: Rodopi. 163-184.
- Izre'el, S. and G. Rahav (2004), 'The Corpus of Spoken Israeli Hebrew (CoSIH); phase I: the pilot study', in: N. Oostdijk, G. Kristoffersen and G. Sampson (eds.) *LREC 2004: fourth international conference on language resources and evaluation; workshop proceedings: compiling and processing spoken language corpora*, Lisbon (Portugal). 1-7.
- Jelinek, F. (1976), 'Continuous speech recognition by statistical methods', *Proceedings of the IEEE*, 64 (4): 532-556.
- Johansson, S. (ed.) (1982), *Computer corpora in English language research*. Bergen: NAVF.
- Kehoe, A. and A. Renouf (2002), 'WebCorp: applying the web to linguistics and linguistics to the web', in: D. Lassner, D. DeRoure and A. Iyengar (eds.) *WWW2002 conference*, Honolulu (Hawaii). New York: ACM. Available from: <http://www2002.org/CDROM/poster/67/>.
- Kurzweil, R. (1990), *The age of intelligent machines*. Cambridge (MA): MIT Press.
- Kytö, M., J. Rudanko and E. Smitterberg (2000), 'Building a bridge between the present and the past: a corpus of 19th-century English', *ICAME journal*, 24: 85-97.
- 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: Renouf, A. and A. Kehoe (eds.) *The changing face of corpus linguistics: papers from the 24th international ICAME conference*, Guernsey, 23-27th April 2003. Amsterdam/New York: Rodopi. 185-204.
- Mair, C. (1997), 'Parallel corpora: A real-time approach to the study of language change in progress', in: M. Ljung (ed.) *Corpus-based studies in English*. Amsterdam/Atlanta: Rodopi. 195-209.
- McCarthy, M. (ed.) (1988), *Special issue on 'Naturalness in language'*, *ELR Journal*.
- Nevalainen, T. (2003), *Historical sociolinguistics*. London: Longman.
- Nevalainen, T. and H. Raumolin-Brunberg (1996), 'The corpus of Early English correspondence', in: T. Nevalainen and H. Raumolin-Brunberg (eds.) *Sociolinguistics and language history. Studies based on the corpus of Early English correspondence*. Amsterdam/Atlanta: Rodopi. 38-54.

- Owen, C. (1988), 'Naturalness and the Language Learner', in: M. McCarthy (ed.) *Special issue on 'Naturalness in language'*, ELR Journal. 21-46.
- Renouf, A. (1987), 'Corpus development', in: J. McH. Sinclair (ed.) *Looking up*. London/Glasgow: Collins ELT. 1-15.
- Renouf, A. (2002), 'The time dimension in modern corpus linguistics', in: B. Kettemann and G. Marko (eds.) *Teaching and learning by doing corpus analysis. Papers from the 4th international conference on teaching and learning corpora*, Graz, 19-24 July 2000. Amsterdam/Atlanta: Rodopi. 27-41.
- Renouf, A. (2003), 'WebCorp: providing a renewable data source for corpus linguists', in: S. Granger and S. Petch-Tyson (eds.) *Extending the scope of corpus-based research: new applications, new challenges*. Amsterdam/Atlanta: Rodopi. 39-58.
- Renouf, A., A. Kehoe and J. Banerjee (2006), 'WebCorp: an integrated system for web text search', in: N. Nesselhauf, M. Hundt and C. Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi. 47-68.
- Rissanen, M. (1992), 'The diachronic corpus as a window to the history of English', in: J. Svartvik (ed.) *Directions in corpus linguistics: proceedings of Nobel Symposium 82*, Stockholm 4-8 August 1991. Berlin: Mouton de Gruyter. 185-205.
- Rissanen, M. (2000), 'The world of English historical corpora', *Journal of English Linguistics*, 28 (1): 7-20.
- Salinger, J. D. (1951), *Catcher in the Rye*. Boston: Little Brown & Co.
- Sinclair, J. McH. (ed.) (1987), *Looking up*. London/Glasgow: Collins ELT.
- Sinclair, J. McH. (1988), 'Naturalness in language', in: M. McCarthy (ed.) *Special issue on 'Naturalness in language'*, ELR Journal. 11-20.
- Summers, D. (1993), 'Longman/Lancaster English language corpus – criteria and design', *International Journal of Lexicography*, 6 (3): 181-208.
- Taavitsainen, I. and P. Pahta (1997), 'The corpus of Early English medical writing', *ICAME Journal*, 21: 71-78.

This page intentionally left blank

Seeing through multilingual corpora

Stig Johansson

University of Oslo

Abstract

In the last 10-15 years there has been a rapidly growing interest in multilingual corpora. In this paper I comment on the development and give some examples from recent research, drawing in particular on work connected with the English-Norwegian Parallel Corpus and the English-Swedish Parallel Corpus. Notions dealt with include translation paradigms, mutual correspondence, semantic mirrors, zero correspondence, and translation effects. Special attention is paid to the English nouns person and thing in a contrastive perspective.

1. Corpora – a way of seeing

The title of this talk is inspired by some words attributed to Henrik Ibsen, the renowned Norwegian author: *Å dikte er å se*. As this is hard to express well in English, I will explain it first in German: *Zu dichten ist zu sehen*. In English we could perhaps say: *Writing is seeing*. It is understood that what has been seen must also be expressed, but first it has to be seen. And so it is with research.

This brings me to corpora. It has often been said that, through corpora, we can observe patterns in language which we were unaware of before or only vaguely glimpsed. My claim is that this applies particularly to multilingual corpora. We can see more clearly what individual languages are alike, what they share and – perhaps eventually – what characterises language in general. Seeing through corpora we can see through language.

2. Parallel texts – old and new

In the last 10-15 years or so there has been a great deal of interest in the development and use of multilingual or parallel corpora. To begin with, we can define such corpora provisionally as collections of texts in two or more languages which are parallel in some way, either by being in a translation relationship or by being comparable in other respects, such as genre, time of publication, intended readership, and so on. I will come back to this point in a moment.

Parallel texts have been used more or less systematically before the age of computers. A famous early example is the Rosetta Stone, discovered by a young French officer in 1799 in Rosetta, a small town near Alexandria. The stone, which is now on display in the British Museum, contains inscriptions in three distinct scripts: Egyptian hieroglyphs, demotic script (a late cursive form of hieroglyphs),

and Greek. A comparison of these texts eventually led to the deciphering of the hieroglyphs.

If we move closer to our own time, we find that parallel texts in different languages have been used both in translation studies and in comparative language studies. Vilém Mathesius, founder of the Linguistic Circle of Prague, spoke about analytical comparison, or linguistic characterology, as a way of determining the characteristics of languages and gaining a deeper insight into their specific features (Mathesius 1975). He used it in his comparison of the word order of English and Czech, and the study was later followed up by Jan Firbas in particular. In the opening chapter of his *Functional sentence perspective in written and spoken communication*, Firbas (1992: 3ff.) compares an original text in French with its translations into English, German, and Czech, and he uses the same sort of comparison later in the book. Firbas writes:

The contrastive method proves to be a useful heuristic tool capable of throwing valuable light on the characteristic features of the languages contrasted. (Firbas 1992: 13)

In a paper from the 1960s we find the notion of translation paradigms. These are forms and their possible translations, with notes on conditions of use. The author suggests that contrastive statements

may be derived from either (a) a bilingual's use of himself as his own informant for both languages, or (b) close comparison of a specific text with its translation. (Levenston 1965: 225)

The use of multilingual corpora, with a variety of texts and a range of translators represented, increases the validity and reliability of the comparison. It can be regarded as the systematic exploitation of the bilingual intuition of translators, as it is reflected in the pairing of source and target language expressions in the corpus texts.

To sum up, we can say that what is new is not the use of parallel texts for language studies, but the fact that multilingual corpora are now being compiled in a systematic manner and are prepared for search and analysis by computer. But what do these corpora look like?

3. Corpus models

So far, I have referred to the comparison of original texts and their translations. But we can distinguish between two main types of multilingual corpora:

- translation corpora: consisting of original texts and their translations into two or more other languages;

- comparable corpora: consisting of original texts in two or more languages matched by criteria such as genre, time of publication, etc.

Both types have their advantages and their problems, as has often been pointed out (see e.g. Johansson 1998). Fortunately, it is not necessary to choose between them. Both types can be combined within the same overall framework, as has been done with the *English-Norwegian Parallel Corpus (ENPC)*; see Figure 1. This is a slightly modified version of a diagram first presented at the ICAME conference in Feusisberg in 1993 (see Johansson and Hofland 1994).¹ The reason for the double arrows is that the comparison can proceed in either direction, e.g. from original text to translation or from translation to the sources in the original text. With this model, which we can call the bidirectional translation model, we can compare:

- original texts across languages;
- original texts and translations across languages;
- original and translated texts within each language;
- translations across languages.

So the corpus changes depending upon our point of view. We have a network of different types of corpora, and each can be used to control and supplement the other. Our claim is that, with this sort of structure, we can learn both about languages and about translation.

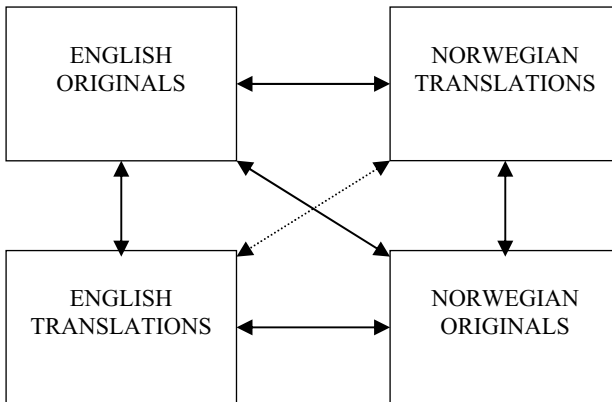


Figure 1: The model for the *ENPC*.

The problem with the bidirectional translation model is that it limits the corpus to texts and text types that have been translated. But all texts are of course not translated, nor are texts and text categories translated to the same extent in both directions. So, corpora of this kind must be supplemented by larger monolingual

corpora to adequately represent the languages to be compared. This is why the original model presented at Feusisberg was as shown in Figure 2.

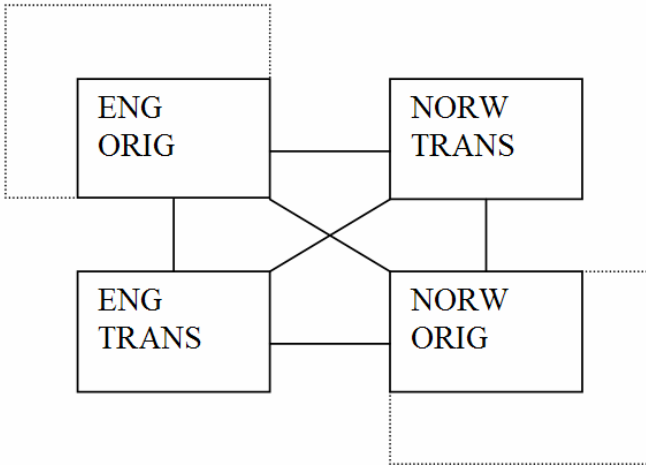


Figure 2: The original model for the *ENPC*.

The problem becomes even more acute if we expand the model to three languages, as we have done at the University of Oslo in a project undertaken in collaboration with the Department of Germanic Studies. We call this the diamond model; see Figure 3 (originally drawn in this manner by my colleague Cathrine Fabricius-Hansen). The number of texts translated across these three languages is limited, and the text types cannot be satisfactorily matched.

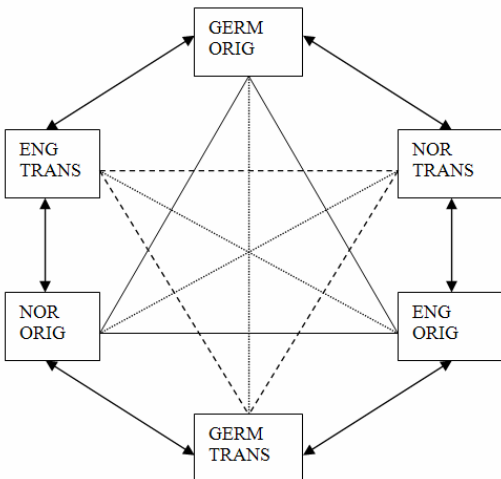


Figure 3: The *Oslo Multilingual Corpus*: English-Norwegian-German.

Because of these problems, it is unlikely that we will ever be able to build up a corpus fully according to the diamond model. But we find it valuable to build this type of resource. The more languages we include, the more clearly can we see the characteristics of each language, and the more general questions can we ask about the nature of language and the characteristics of translation. Apart from English, Norwegian, and German, we also have texts in other languages forming part of the *Oslo Multilingual Corpus*.²

There is a good deal more to say about corpus models. A great deal can also be said about the development of tools for building and using multilingual corpora, such as aligners and browsers. Above all, much could be said about multilingual corpus projects in different parts of the world. In February 2004 I made a search on the Internet and found over 2,300 references to the term multilingual corpora. There are both more linguistically oriented projects and projects concerned with automatic language processing tasks, such as machine translation, multilingual terminology extraction, and multilingual information retrieval. As it is impossible to give an adequate account of all the various projects,³ I have chosen to limit my discussion to some examples from the work I know best, viz. that connected with the *ENPC* and its Swedish sister project, the *English-Swedish Parallel Corpus (ESPC)*.⁴ I turn now to the most important question I want to pose: What is it we can observe through multilingual corpora?

4. What can we see?

One of the things which have surprised me most over the years is how reluctant many linguists have been to look into corpus evidence. But the tide has turned, and more and more linguists are now willing to see. It is a common experience among those who work with corpora that we often make new discoveries. One of the authors of the *Cambridge grammar of the English language* (Huddleston and Pullum 2002), a grammar where corpus use is not particularly obtrusive, recently put it in this way:

I have said this before and I will say it again: the grammarian who looks into a corpus usually learns at least one or two new things from the first relevant sentence, other than the thing that was initially being investigated.

(Pullum, quoted from: <http://itre.cis.upenn.edu/~myl/linguagelog/archives/000198.html>)

What really counts in linguistics, as in research in general, and indeed in life itself, is having an open mind.

When we turn to a multilingual corpus, we will quickly find that what we may initially have regarded as alike may turn out to differ in important respects. The comparison sharpens the perception of both similarities and differences. In his thesis on presentative constructions in English and Norwegian, based on the

ENPC, Jarle Ebeling (2000) studies three constructions which are found in both languages, termed full presentatives (1), bare presentatives (2), and *have/ha-*presentatives (3):

- (1) There's a long trip ahead of us.
Det ligger en lang reise foran oss.
- (2) A long trip is ahead of us.
En lang reise ligger foran oss.
- (3) We have a long trip ahead of us.
Vi har en lang reise foran oss.

Although the constructions in the two languages are similar in syntax and semantics, there are important differences in use. The same applies to cleft constructions in English and Swedish, investigated on the basis of the *ESPC* by Mats Johansson (2002). The contrastive study defines the differences and at the same time makes the description of the individual languages more precise.

If we turn to lexis, we rarely find full equivalence across languages. In a number of studies Åke Viberg has compared Swedish and English verbs on the basis of the *ESPC*, for example Swedish *gå* and English *go* (Viberg 1996), Swedish *få* and English *get* (Viberg 2002). Although these verbs share a core meaning and show many similarities in meaning extension, differences are surprisingly large. The cross-linguistic perspective brings out the diverging polysemy in the two languages. At the same time, the study reveals general tendencies in semantic development.

5. Some descriptive notions

My next task is to briefly present some notions which we have found useful in working with our corpora. For illustration I draw chiefly on examples from the *ESPC* and the *ENPC*. As I cannot take any knowledge of Scandinavian languages for granted, I have chosen examples which may throw light on English, the language which all of us here have in common. The question asked is: What does English look like from a contrastive perspective?

5.1 Translation paradigms

The first notion I want to introduce is 'translation paradigms', a term which is generally used about different approaches to translation. In our context we mean the forms in the target text which are found to correspond to particular words or constructions in the source text.⁵ As an example, consider the recent paper by Aijmer and Simon-Vandenberg (2003) on the English discourse particle *well*, based on the *ESPC* and the English-Dutch texts of the *Oslo Multilingual Corpus*.

The main object of the study was to examine the function of *well* in the light of its Swedish and Dutch translations.

In both languages the translators have drawn on many different forms: discourse particles, modal particles, combinations of discourse particles and modal particles, conjunctions, concessive adverbials, etc. These make up the translation paradigms, and the task of the analyst is to analyse their meanings and conditions of use. The study shows the multifunctionality of *well*, which the authors claim is derivable from a more general function: “to turn the utterance into a heteroglossic one, signalling the speaker’s awareness of the heterogeneity of views, positioning the utterance in the context of preceding and following texts.” (Aijmer and Simon-Vandenberg 2003: 1155)

5.2 Mutual correspondence

Using a bidirectional corpus, we can show correspondences in both directions, i.e. sources as well as translations. To continue with the example above, we can ask both ‘How is *well* translated?’ and ‘Where does *well* in English translated texts come from?’, as I have done in a recent study (Johansson 2006).

With a bidirectional corpus, we can also calculate mutual correspondence, i.e. the degree of cross-linguistic correspondence between particular forms or categories of forms, as shown by Bengt Altenberg (1999) in a study of adverbial connectors in English and Swedish. The mutual correspondence measure varies between 0% (= no correspondence) and 100% (= full correspondence). In practice, there is hardly ever full correspondence, but varying degrees depending upon the forms or form groups compared. For example, Altenberg finds that listing conjuncts correspond in approximately 80% of the cases, whereas the figure is considerably lower for inferential and transitional conjuncts (and particularly for explanatory conjuncts). Correspondences can be asymmetric and differ depending upon the direction of translation (this is termed translation bias), as with the two contrastive conjuncts *however* and Swedish *emellertid*. Going from *emellertid* to English the correspondence figure is as high as 81%, i.e. four out of five cases are translated by *however*. Starting from *however* the figure is only 47%, i.e. less than half of the cases are translated by *emellertid*. Working in this way, Altenberg eventually arrives at a cross-linguistic comparison of entire subsystems in the two languages.

5.3 Semantic mirrors

In monolingual corpora we can easily study forms and formal patterns, but meanings are less accessible. One of the most fascinating aspects of multilingual corpora is that they can make meanings visible through translation. Ambiguity and vagueness are revealed through translation patterns. To be sure, the reflections are not perfect. As always, the corpus user must approach the material with care. Corpus use does not eliminate assessment of the evidence.

The use of language comparison to reveal meanings seems to be implicit in the title of a recent book: *Meaning through language contrast* (Jaszczolt and

Turner 2003). In a recent study Dirk Noël shows how “translators, through the linguistic choices they make, inadvertently supply evidence of the meanings of the forms they are receiving and producing” (Noël 2003: 757). On the basis of translation correspondences in the Canadian Hansard corpus (English-French), he shows that forms like *BE said to* and *BE reported to* are turning into evidential auxiliaries.

Multilingual corpora are also exploited consistently to reveal meaning in Viberg’s lexical studies and in Aijmer and Simon-Vandenberg’s investigation of *well* in a cross-linguistic perspective. But the most radical attempt to draw semantic information from a multilingual corpus is Helge Dyvik’s (1998) semantic mirrors project. Taking the *ENPC* as a starting-point, Dyvik examines correspondences of words first in one direction (first image), then from each of these back to the first language again (second image), and a third time from each of the new correspondences across to the other language (third image). In the criss-crossing between the languages, the number of words grows dramatically, resulting in a complex word net. The ultimate aim of the project is to lay a new foundation for semantics.

5.4 Zero correspondence

It is a common experience in using translation corpora that there may be no formal cross-linguistic correspondence. We call this omission, or zero correspondence. In their study of *well* Aijmer and Simon-Vandenberg found 21% zero correspondence in the English-Swedish material and 7% in the English-Dutch material. In his work on adverbial connectors, Altenberg found 30% omission for inferential conjuncts (e.g. *in that case, then, otherwise*) and 34% omission for transitional conjuncts (e.g. *incidentally, now, by the way*).

Zero correspondence is not just a matter of omission; often there is addition, without any formal counterpart in the original text. In my study of *well* I found that there was 16% zero correspondence in going from English to Norwegian and 21% in the other direction. In other words, about every fifth occurrence of *well* in the English translations had no clearly identifiable source. The reason for the omission in translation is often that there is some kind of compensation, as in (4) and (5):⁶

- (4) “But you have at some point to understand that your father is not prepared any longer to share his ill-gotten gains with Jasper and all his friends.”
 “*Well, at least* he is prepared to see they are ill-gotten,” said Alice earnestly. (DL2, orig)
 “Men før eller senere vil du likevel bli nødt til å innse at din far ikke lenger er innstilt på å dele sine urettmessig ervervede rikdommer med Jasper og alle vennene hans.”
 “Han er *i det minste* i stand til å innrømme at de er urettmessig ervervet,” sa Alice alvorlig.

- (5) Sonia wasn't daft. *Well, not then, anyway*, unless that's what the abandoning of your own life for your children can be called. (FW1, orig)
 Sonia var ingen tosk. *Ikke da i hvert fall*, bortsett fra hva det at du kaster bort livet ditt for barnas skyld, kan bli kalt.

Here the qualification is conveyed by *i det minste* ('at least') and *i hvert fall* ('anyway'), and *well* can be dispensed with in the translation. Similar examples were found in the material examined by Aijmer and Simon-Vandenberg.

In other cases, it is impossible to find a good reason for zero correspondence, and we may assume that translators have responded to the whole situation reported in the text. In the next two examples *well* has been added by the translator, although there is no formal counterpart in the source text:

- (6) "Jaså, forsvunnet fra et aldershjem?
 Hvem meldte henne savnet?
 Javel.
 Vi snakkes ved." (EG1, orig)
 "Reported missing from an old people's home?
 Who reported it?
 Oh?
Well, we'll talk about that later."
- (7) "Har dere tatt knekkebrød igjen?"
 Hildegun himlet lidende mot taket og svarte med uforskammet høflighet:
 "Neida, mor. Vi drikker bare nypete."
 "Den koster også penger. Dere har brukt strøm." (BV1, orig)
 "Have you been at the crisp-bread again?"
 Hildegun rolled her eyes in suffering towards the ceiling and answered with brazen politeness.
 "No, mother, we haven't. We're just drinking rose-hip tea."
 "*Well* that costs money too. You've been using electricity."

In (6) *well* is used to round off a telephone conversation. Example (7) opens with a reproach in the form of a question; the child tries to take the brunt of the reproach by using *just* (corresponding to *bare* in the Norwegian original); the mother continues with another critical remark, opening with a mitigating *well*, without a counterpart in the original. The need for such grease in the interaction apparently varies in different languages, as do the means of expression.

Although zero correspondence may be the result of carelessness on the part of the translator or may be due to conscious adaptation of the text to the target audience, it is chiefly interesting for the linguist in that it highlights cross-linguistic differences. In a well-known paper on translation Roman Jakobson describes language differences in this way: "Languages differ essentially in what they *must* convey and not in what they *may* convey." (Jakobson 1966: 236). I would rather put it in this way: languages differ not so much in what they may

convey as in what they conventionally convey in natural discourse. One of the most important aspects of multilingual corpora is that they make it possible to go beyond a comparison of language systems *in abstracto* to a study of languages in use.

5.5 Parallel translations

So far I have been concerned with a straightforward comparison going either from original texts to their translations, or from translations to their sources in original texts. Given a multilingual corpus we can also study parallel translations, as I have done in a comparison of English *that's what* constructions and their correspondences in Norwegian and German (Johansson 2001) and in a study of generic subject pronouns in English, Norwegian and German (Johansson 2004a).

To illustrate the notion of parallel translations, I will pick a couple of examples from a subcorpus of the Oslo Multilingual Corpus containing Norwegian original texts and their translations into three other languages; see Figure 4. Here we can of course make a direct comparison of how features of the Norwegian original texts are conveyed in the other languages, but we can also make an indirect comparison starting from English, French or German.

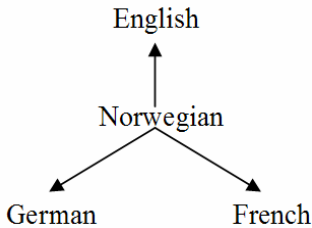


Figure 4: The *Oslo Multilingual Corpus*: Norwegian-English-German-French.

Suppose I am interested in comparing English supplementary *-ing* clauses with their correspondences in the other languages. If I search for such structures, I will find their sources in the Norwegian original text, and I will also see what forms have been chosen by the German and the French translators. The most typical pattern is shown in (8) and (9):

- (8) “Ssssh,” the boy whispers, *placing* a finger to his lips.
 – Hyss, hvisker gutten *og legger* [lit. ‘and places’] fingeren på munnen.
 (BHH1, orig)
 “Pst”, flüstert der Junge *und legt* den Finger an die Lippen.
 – Chut! murmura le garçon *en se mettant* un doigt sur la bouche.
- (9) For a while they stood motionless, *staring* up the avenue as if unable to believe their eyes.

En liten stund sto de urørlige *og kikket* [lit. ‘and looked’] oppover alleen som om de hadde vondt for å tro sine egne øyne. (BHH1, orig)

Eine Weile standen sie dann bewegungslos da *und blickten* die Allee entlang, als trauten sie ihren eigenen Augen nicht.

Pendant un moment, elles restèrent ainsi immobiles, *regardant* l’allée comme si elles n’en croyaient pas leurs yeux.

The Norwegian original and the German translation have coordination, while the English and French translators opted for participle constructions of the type which is called *participium coniunctum* in Latin grammar. French and English do of course have structures of coordination, but the translators have preferred to restructure the sentence. As this is a regular pattern, though by no means the only one, there must be an explanation.

Turning to Mustanoja’s Middle English grammar, I found the following comment on participle constructions:

The appositive use of the present and past participles (the *participium coniunctum* of the Latin grammarians) is a well-known feature in many languages. It is common in classical Latin, even more common in the Vulgate, and profuse in medieval Latin. [...] Some aspects of this use seem to be native in origin, such as the adjectival use of the past participle (i.e., its use as an equivalent of a relative clause), while the corresponding use of the present participle and the use of the participles as equivalents of various adverb clauses and of co-ordinate clauses are evidently due to Latin influence. Usually, however, Latin appositive participles are rendered into OE by means of co-ordinated finite verbs or somewhat less frequently of subordinated finite verbs. (Mustanoja 1960: 554-555)

German and Norwegian evidently preserve the old Germanic pattern, with coordination. English has adopted the Romance participle construction, and it has become a regular feature of the language, presumably because it has been found useful in making it possible to distinguish between simultaneous events, marked by *-ing* constructions, and consecutive events, marked by coordination. German and Norwegian, however, use coordination for both purposes.⁷

5.6 Translation effects

Above I have shown how translations can be used to reveal characteristics of different languages. But translation corpora must be used with caution in cross-linguistic research.⁸ For one thing, translations may reflect features of the source language, a phenomenon which has been given the label *translationese* (see e.g. Gellerstam 1996). Moreover, there may be general features characteristic of translated texts (see e.g. Baker 1993). The problem can be turned to an advantage with the *ENPC* model, where it is possible to control for translation effects.

We can pinpoint differences in distribution between elements in original and translated texts in the same language. For example, the discourse particle *well* is far less common in English texts translated from Norwegian than in English original texts (Johansson 2006), presumably because Norwegian lacks a fully-fledged counterpart of this form. Supplementive *-ing* clauses are less common in texts translated from Norwegian than in English original texts (Westre 2003). Even such basic features as word order and tense choice may be influenced by the source text (see Hasselgård 2000 and Elsness 2000/2001).

These results suggest that translators have a tendency to move on the surface of discourse. But we should be careful in using the term *translationese*, which – like other words ending in *-ese* – has a negative ring. Translation effects are not necessarily negative, because translation is an important way of renewing the target language. The degree of source vs. target orientation may vary depending upon a number of factors, such as text category or the degree of importance of the source text.

My claim is that corpora compiled according to the bidirectional translation model can both exploit translation as an instrument in cross-linguistic research and be used to pinpoint characteristics of translated texts.

6. Two case studies

As additional examples of the use of multilingual corpora, I will briefly examine two common English nouns, *person* and *thing*, from the point of view of Norwegian, again on the basis of the *ENPC*. To speak with Halliday and Hasan (1976: 274), both are general nouns on the borderline between a lexical and a grammatical item. In both cases we have cognate words in the two languages, derived from the same sources, and at the outset we might not have expected much of a difference.

6.1 English vs. Norwegian *person*

English and Norwegian *person*⁹ both go back to the Latin word *persōna*, originally meaning a mask worn by an actor or a part played by an actor. Judging by the entries in dictionaries for present-day English and Norwegian, they appear to have much the same meaning and use. But the overall distribution in the corpus is intriguing; see Table 1. As the English and the Norwegian parts of the corpus are balanced, we can compare raw frequency figures. What do we see?

- In the first place, the English word is much more common than its Norwegian counterpart.
- In the second place, the frequency is strikingly high in translations from Norwegian. This is shown more clearly in Figure 5. The same pattern is found in the *ESPC*, as shown in Figure 6.

Table 1: The overall distribution of *person* in the ENPC.¹⁰

	<i>fiction</i>		<i>non-fiction</i>	
	<i>original</i>	<i>translation</i>	<i>original</i>	<i>translation</i>
N <i>person(en)</i>	25	35	31	20
E <i>person</i>	67	94	39	110

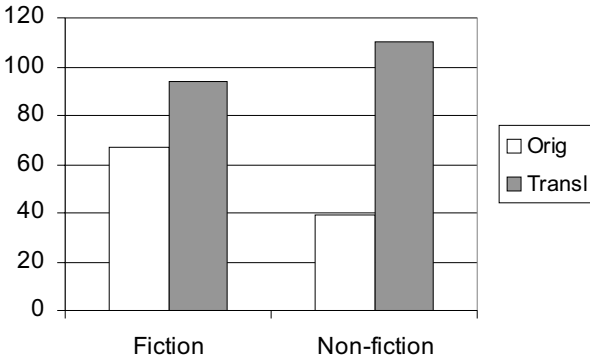


Figure 5: The overall distribution of English *person* in the ENPC.

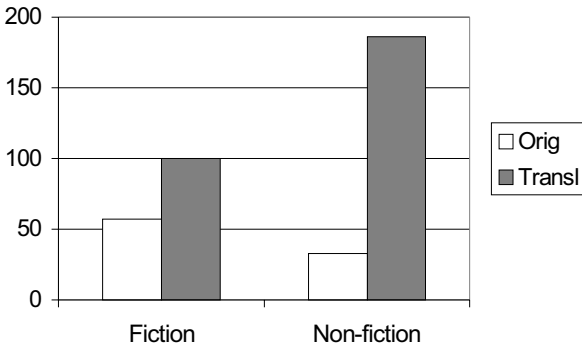


Figure 6: The overall distribution of English *person* in the ESPC.

What might the reason be for these differences in distribution? Apparently English *person* has a wider area of use than its Norwegian cognate. The reason is partly that it corresponds both to *menneske* ('human being')¹¹ and *person* in Norwegian. Equally important, English *person* seems to be used as a kind of pronoun in cases where Norwegian can manage without a noun, as in:

- (10) Big exciting news that *the first person* didn't recognise any more.
Store spennende ord som *den første* [lit. 'the first'] ikke kjente igjen.
(THA1, orig)
- (11) He walks one step behind me, I've never led *a blind person* before.
Han går et skritt bak meg, jeg har aldri ført *en blind* [lit. 'a blind'] før.
(CL1, orig)
- (12) Members may be reappointed, but *no person* may serve on the Executive Board for more than two consecutive terms.
Medlemmene kan gjenoppnevnes, men *ingen* [lit. 'nobody'] kan sitte sammenhengende i hovedstyret i mer enn to perioder. (NFRV1, orig)
- (13) The essential point is for *the person* actually caring for the child to be able to take the leave.
Det avgjørende er at *den* som [lit. 'the/that who'] faktisk har omsorgen for barnet, kan ta permisjonen. (S11, orig)

In (10) and (11) Norwegian has a nominalised adjective where English requires a head noun. In (12) the English translator has inserted a noun phrase headed by *person*, although it would have been possible to use a pronoun like *nobody*. In (13) *person* is inserted to support the following relative clause. A different type of insertion is found in (14), where a noun phrase headed by *person* is used to convey Norwegian compounds without a straightforward counterpart in English:

- (14) There's a great difference in the underwear of *a person from the north in the winter and the same person in the summer*.
Det er stor forskjell på underklærne til *en vinternordlending og en sommernordlending* [lit. 'a winter northerner and a summer northerner'].
(HW1, orig)

English *person* seems to have a lexicogrammatical function which I have tried to capture by the term *pronoun*. The *pronoun* is used to provide a head to support elements which cannot on their own fill a nominal position. Such uses may eventually give rise to new pronouns. It is interesting to note the parallel with the historical development leading to the English pronouns ending in *-body*.¹² The French counterpart *personne* has made it all the way to a pronoun.

This is all I can say in this context about *person* and its correspondences in Norwegian. The main point is that the lexicogrammatical function of the English word is brought to light through the cross-linguistic perspective. For a more detailed account, I refer to a recent paper (Johansson 2004b). My next example takes up another common noun which also exhibits intriguing cross-linguistic correspondences.

6.2 English *thing* vs. Norwegian *ting*

English *thing* and Norwegian *ting*¹³ have a common Germanic origin, and at the outset the two words would seem to be used in much the same way in present-day English and Norwegian. What do we see if we consult the *ENPC*? The overview in Table 2 shows that, as with *person*, overall frequencies are higher in English, particularly so in fiction where there are more than twice as many occurrences as for the Norwegian word. There is also a translation effect, though not as marked as for *person*. In this case I will focus on the degree to which the two words correspond, starting from English *thing*; see Figures 7 and 8.

Table 2: The overall distribution of English *thing* and Norwegian *ting* in the *ENPC*.¹⁴

	<i>fiction</i>		<i>non-fiction</i>	
	<i>original</i>	<i>translation</i>	<i>original</i>	<i>translation</i>
N <i>ting</i>	224	253	113	79
E <i>thing</i>	499	567	123	147

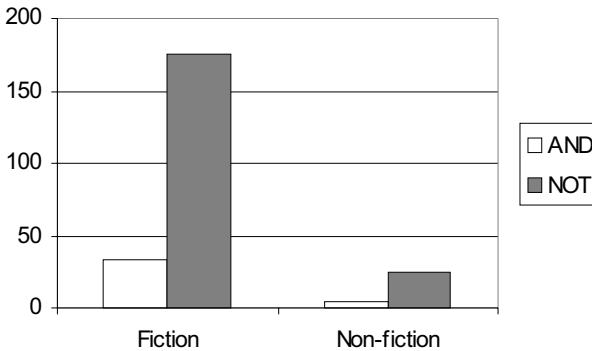


Figure 7: English *thing* vs. Norwegian *ting* (singular or plural), based on English original texts in the *ENPC*; AND = the forms correspond, NOT = the forms do not correspond.

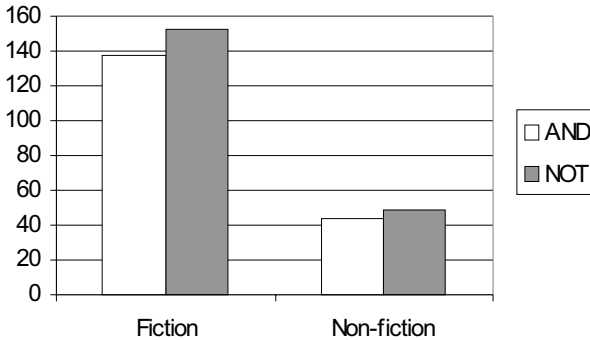


Figure 8: English *things* vs. Norwegian *ting* (singular or plural), based on English original texts in the ENPC); AND = the forms correspond, NOT = the forms do not correspond.

What can we see?

- In the first place, there is a lot of non-correspondence, as we would expect, for the simple reason that the Norwegian word is less frequent overall than its English cognate.
- More interestingly, there is a striking difference in the degree of non-correspondence for the singular form *thing* and the plural form *things*. With singular *thing*, the great majority of the instances do not correspond to Norwegian *ting*.

What might be the explanation for the non-correspondence in the case of the singular form? If we look into the corpus material, we quickly find a clue. Some of the instances are similar to those found for *person*, as in:

- (15) That seemed to be *the appropriate thing* to do. (AB1, orig)
 Det virket som *det beste* [lit. ‘the best’] hun kunne gjøre.
- (16) “How could you do *such a thing*?” she asked Edward. (AT1, orig)
 “Åssen kunne du gjøre *noe sånt* [lit. ‘some such’]?” spurte hun Edward.

Above all, we find many more or less fixed sequences: *the first thing*, *the last thing*, *the only thing*, *the same thing*, *kind of thing*, *sort of thing*, *the whole thing*, *a thing or two*, *another thing*, etc.¹⁵ It is these collocational patterns which lead to the great differences in relation to the Norwegian cognate *ting*. Again the characteristics of the English noun are illuminated through the cross-linguistic perspective. But the thing is there is no time now to go further into these things.

7. Problems and prospects

It is time for a summing up and for trying to look ahead. In my paper I have mainly referred to Scandinavian corpus studies, using our bidirectional translation model. But as I have pointed out, this model is not without its problems. We need other models. Above all, we need to make multilingual corpus studies truly multilingual. This is an important task for the future. The more languages we include, the more general questions can we ask about language and translation.

Other important tasks for the future are to try to unite multilingual corpus studies of the more linguistic and the more computational kind; cf. Lars Borin's discussion in his paper "... and never the twin shall meet" (in Borin 2002: 1-43). We also need to connect contrastive studies of the more theoretical and the more pedagogical kind. An important development is the Integrated Contrastive Model which combines learner corpora and multilingual corpora (Granger 1996, 2000/2001). More fundamentally, we need to recognise that the corpus is only one way of seeing language. As shown by von Stutterheim et al. (2002), we can gain insight into language and conceptualisation by cross-linguistic experiments. Finally, we must never forget the crucial role of intuition in all language studies.

Recall how we started with Ibsen's saying *Å dikte er å se*. The Norwegian professor of literature Francis Bull adds: *men man må se energisk* ("but one must look energetically").¹⁶ If we are prepared to look energetically into multilingual corpora, we can see correspondences across languages, we can see individual languages in a new light, we can pinpoint characteristics of translation, we can see meanings, we can see grammaticalisation, we can see collocations, we can see the intimate relationship between lexis and grammar. Seeing through corpora we can see through language.

What we have seen so far is promising. But there is far more to discover, because the exploration of languages through multilingual corpora has only just begun.

Notes

- 1 For more information on the *English-Norwegian Parallel Corpus*, see: <http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/>
- 2 For information on the *Oslo Multilingual Corpus*, which includes texts in a number of languages, see: <http://www.hf.uio.no/ilos/OMC/>.
- 3 Some relevant references are: Borin (2002), Botley et al. (2000), and Véronis (2000).
- 4 For information on the *English-Swedish Parallel Corpus (ESPC)*, see: <http://www.englund.lu.se/research/corpus/corpus/esp.html>.

- 5 Cf. the paper by Levenston (1965) referred to in Section 2 above.
- 6 The examples given here, and later in the paper, are from the *ENPC*. For an explanation of the abbreviated references, see the ENPC website (note 1)
- 7 Consulting one of the Latin grammars written for Swedes, I found a recommendation from the author to make the logical relationship clear by an adverb (meaning ‘thereby’, ‘thereafter’, ‘therefore’, etc.) where coordination is used to translate *participium coniunctum* (Sjövall 1953: 322).
- 8 See e.g. Lauridsen (1996) and Teubert (1996).
- 9 This topic was studied in a term paper by Eli Grinde in connection with my course on ‘Contrastive analysis and learner language’ (Grinde 2003).
- 10 Only singular forms are included here, for Norwegian both the indefinite form *person* and the definite form *personen*.
- 11 Originally a nominalised adjective derived from proto-Germanic **mannisko-*, from **mann* + *-isko-* (‘human’). Cf. Swedish *människa*, German *Mensch*, Old English *mennisc*.
- 12 Cf. *OED*, *body*, 13: A human being of either sex, an individual. Formerly, as still dialectally, and in the combinations ANY-, EVERY-, NO-, SOMEBODY, etc., exactly equivalent to the current ‘person’; but now only as a term of familiarity, with a tinge of compassion, and generally with adjectives implying this.
- 13 This topic was studied in an unpublished term paper by Bohumila Chocholousová (2003).
- 14 Both singular and plural forms are included here, more exactly all forms starting with *thing* for English and all forms starting with *ting* for Norwegian. The reason for combining the singular and the plural forms is that Norwegian *ting* takes no plural inflection. The Norwegian figures include some instances of *ting* used with reference to a legal or political assembly, but these do not materially affect the findings.
- 15 For more details, see the paper by Chocholousová (2003).
- 16 See Bull (1966: 34).

References

- Aijmer, K. and A.-M. Simon-Vandenberg (2003), 'The discourse particle *well* and its equivalents in Swedish and Dutch', *Linguistics*, 41: 1123-1161.
- Altenberg, B. (1999), 'Adverbial connectors in English and Swedish: semantic and lexical correspondences', in: H. Hasselgård and S. Oksefjell (eds.) *Out of corpora. Studies in honour of Stig Johansson*. Amsterdam: Rodopi. 249-268.
- Baker, M. (1993), 'Corpus linguistics and translation studies: implications and applications', in: M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and technology. In honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins. 233-250.
- Borin, L. (ed.) (2002), *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April 1999*. Amsterdam: Rodopi.
- Botley, S. P., A. M. McEnery and A. Wilson (eds.) (2000), *Multilingual corpora in teaching and research*. Amsterdam: Rodopi.
- Bull, F. (1966), 'Å dikte er å se', in: *Vildanden og andre essays*. Oslo: Gyldendal. 29-37.
- Chocholousová, B. (2003), 'Is the Norwegian *ting* the same *thing* in English: contrasting two words in English and Norwegian', unpublished term paper, Department of British and American Studies, University of Oslo.
- Dyvik, H. (1998), 'A translational basis for semantics', in: S. Johansson and S. Oksefjell (eds.) *Corpora and cross-linguistic research: theory, method, and case studies*. Amsterdam: Rodopi. 51-86.
- Ebeling, J. (2000), *Presentative constructions in English and Norwegian. A corpus-based contrastive study*. Acta Humaniora 60. Oslo: Unipub forlag.
- Elsness, J. (2000/2001), 'A contrastive look at the present/preterite opposition in English and Norwegian', *Languages in contrast*, 3 (2000/2001): 3-40.
- Firbas, J. (1992), *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge University Press.
- Gellerstam, M. (1996), 'Translations as a source for cross-linguistic studies', in: K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies. Lund 4-5 March 1994*, Lund Studies in English 88. Lund: Lund University Press. 53-62.
- Gilquin, G. (2000/2001), 'The integrated contrastive model: spicing up your data', *Languages in contrast*, 3 (2000/2001): 95-123.
- Granger, S. (1996), 'From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora', in: K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies. Lund 4-5 March 1994*, Lund Studies in English 88. Lund: Lund University Press. 37-51.

- Grinde, E. (2003), A person vs. ein person. Are the English *person* and the Norwegian *ein person* the same person?. Unpublished term paper, Department of British and American Studies, University of Oslo.
- Halliday, M. A. K. and R. Hasan (1976), *Cohesion in English*. London: Longman.
- Hasselgård, H. (2000), 'English multiple themes in translation', in: A. Klinge (ed.) *Contrastive studies in semantics*. Copenhagen Studies in Language 25. Frederiksberg: Samfundslitteratur. 11-38.
- Huddleston, R. and G. K. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Jakobson, R. (1966) [1959], 'On linguistic aspects of translation', in: R. A. Brower (ed.) *On translation*. New York: Oxford University Press. 232-239.
- Jaszczolt, K. M. and K. Turner (eds.) (2003), *Meaning through language contrast*. Amsterdam/Philadelphia: Benjamins.
- Johansson, M. (2002), *Clefts in English and Swedish. A contrastive study of IT-clefts and WH-clefts in original texts and translations*. PhD dissertation, Lund University.
- Johansson, S. (1998), 'On the role of corpora in cross-linguistic research', in: S. Johansson and S. Oksefjell (eds.) *Corpora and crosslinguistic research: theory, method, and case studies*. Amsterdam: Rodopi. 3-24.
- Johansson S. (2001), 'The German and Norwegian correspondences to the English construction type *that's what*', *Linguistics*, 39: 583-605.
- Johansson, S. (2004a), 'Viewing languages through multilingual corpora, with special reference to the generic person in English, German, and Norwegian', *Languages in Contrast*, 4 (2002/2003): 261-280.
- Johansson, S. (2004b), 'What is a person in English and Norwegian?', in: K. Aijmer and H. Hasselgård (eds.) *Translation and corpora*. Gothenburg: Acta Universitatis Gothoburgensis. 71-85.
- Johansson, S. (2006), 'How well can *well* be translated? On the discourse particle *well* and its correspondences in Norwegian and German', in: K. Aijmer and A.-M. Simon-Vandenberg (eds.) *Pragmatic markers in contrast*. Oxford: Elsevier. 115-137.
- Johansson, S. and K. Hofland (1994), 'Towards an English-Norwegian parallel corpus', in: U. Fries, G. Tottie and P. Schneider (eds.) *Creating and using English language corpora*. Amsterdam: Rodopi. 25-37.
- Lauridsen, K. (1996), 'Text corpora in contrastive linguistics: which type of corpus for which type of analysis?', in: K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies. Lund 4-5 March 1994*. Lund Studies in English 88. Lund: Lund University Press. 63-71.
- Levenston, E. A. (1965), 'The "translation paradigm": a technique for contrastive syntax', *International review of applied linguistics*, 3: 221-225.
- Mathesius, V. (1975), *A functional analysis of present-day English on a general linguistic basis*, transl. L. Dusková, ed. J. Vachek. Prague: Academia.

- Mustanoja, T. (1960), *A Middle English syntax. Part I: parts of speech*. Mémoires de la Société Néophilologique de Helsinki XXIII. Helsinki: Société Néophilologique.
- Noël, D. (2003), 'Translations as evidence for semantics: an illustration', *Linguistics*, 41: 757-785.
- Sjövall, N. (1953), *Ny latinsk grammatikk*. Lund: Gleerup.
- Teubert, W. (1996), 'Comparable or parallel corpora?', *International journal of lexicography*, 9: 238-264.
- Véronis, J. (ed.) (2000), *Parallel text processing. Alignment and use of translation corpora*. Dordrecht/Boston/London: Kluwer.
- Viberg, Å. (1996), 'Cross-linguistic lexicology. The case of English *go* and Swedish *gå*', in: K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies*. Lund 4-5 March 1994. Lund Studies in English 88. Lund: Lund University Press. 151-182.
- Viberg, Å. (2002), 'Polysemy and disambiguation cues across languages. The case of Swedish *få* and English *get*', in: B. Altenberg and S. Granger (eds.) *Lexis in contrast*. Amsterdam/Philadelphia: Benjamins. 119-150.
- von Stutterheim, C., R. Nüse and J. M. Serra (2002), 'Cross-linguistic differences in the conceptualisation of events', in: H. Hasselgård, S. Johansson, B. Behrens and C. Fabricius-Hansen (eds.) *Information structure in a cross-linguistic perspective*. Amsterdam: Rodopi. 179-198.
- Westre, A. (2003), Supplementive clauses and their correspondences in English and Norwegian fiction texts, unpublished term paper, Department of British and American Studies, University of Oslo.

This page intentionally left blank

Corpora and spoken discourse

Anne Wichmann

University of Central Lancashire, Preston

Abstract

The focus of this paper is on spoken corpora – corpora of naturally occurring speech data that have been compiled for the use of linguists and discourse analysts, as opposed to speech corpora, as commonly used for applications in speech technology and containing various forms of elicited data. I discuss some of the practical and theoretical issues involved in compiling and analysing such data, especially the problems of prosodic annotation and the automatic analysis of the speech signal. I argue that the primary data, i.e. the sound files, are of crucial importance: sounds are not just an additional resource for the study of prosody but an integral part of the message.

1. Introduction

There are many different reasons for studying speech, and there are consequently many different notions of what constitutes a ‘corpus’ of speech. For some it is simply as much data as the analyst can find and analyse, for others it is a carefully balanced and representative sample of different kinds of naturally-occurring spoken language, while for others again it is a large collection of controlled utterances, elicited and often recorded under laboratory conditions. Material used by the speech community is generally gathered with a specific application in mind: speech synthesis, automatic speech recognition, or human-machine dialogue, which involves both of these. The application frequently drives the data, as in collections of service-encounter dialogues which are examined with a view to replicating human patterns in an artificial environment. Similarly, speech recorded under specific conditions, such as sleep deprivation or stress, or in a noisy environment, may be analysed in order to train automatic speech recognition (ASR) software to filter out extraneous and hence potentially distracting aspects of the signal. Material gathered to support research in speech synthesis, on the other hand, may need to be recorded under conditions that eliminate from the start any potentially irrelevant noise. In many cases, naturally-occurring speech is unsuitable for such research because it is not controlled. Use is therefore made of elicited speech – sentences or digit strings read aloud, or dialogue elicited in such a way as to be ‘natural’ in that it is unscripted, but unnatural in that it is highly constrained in subject matter or function, and would not have occurred outside the research context.

For studies of phonetics and phonology, in the field of research known as laboratory phonology, similarly constrained and controlled data may be required for the investigation of underlying phonological systems (research which may or

may not have technological applications). This is based on the assumption that these systems can be identified most easily when not obscured by the features of performance, e.g. disfluencies such as repetition and repair. This, of course, reflects a particular view of language itself, a view very different from the functional approach, which sees language systems as socially constructed and therefore to be sought exclusively in the sequential pattern of natural conversation. Those who take this view, notably Conversation Analysts, while often using data from larger corpora, are committed to minute qualitative analysis of short stretches of conversation, which limits the amount of data and provides results from which it can be hard to generalise.

In the corpus linguistics community, the focus is also on naturally-occurring speech, collected with a view to understanding how human beings communicate verbally rather than how human speech can be simulated for applications in technology. However, the goal is generally to achieve the insights to be gained from large-scale quantitative analysis, which often reveals patterns that case studies or introspection would not provide, and allows generalisations to be made. My claim here is that corpus-based work on spoken language cannot ignore the information contained in the speech signal, whether it is analysed auditorily or acoustically, and whether it is analysed by the individual user or analysed in advance and made available as a phonetic/prosodic transcription.

2. Spoken corpora and prosody

Leech's (2000) overview of spoken English corpora includes only two that have prosodically transcribed: the *London Lund-Corpus (LLC)* (Svartvik 1990) and the *Spoken English Corpus (SEC)* (Knowles et al. 1996). These corpora have, however, given rise to a wide range of discourse studies using a corpus approach, both in relation to conversational interaction and in relation to text structure. Much of this work (e.g. Aijmer 1996, Stenström 1994) has treated transcriptions as primary data, especially work based on the transcribed version of the *LLC*. Although the sound files exist, they were not made publicly available, and in any case, the detailed prosodic analysis was the work of prosody experts, on whose work those focusing on discourse and pragmatics as their main area of expertise were happy to rely.

For the *SEC*, the only other corpus of British English to have been fully transcribed prosodically, the sound files are, by contrast, publicly available, and some work (e.g. Knowles et al. 1996, Wichmann 2000) was thus able to compare the expert transcription with the original sound.¹ This enabled not only the investigation of the phonetics-phonology interface but also provided a basis for studies going beyond what was captured by the transcription. Most of the *SEC* texts are examples of highly prepared or read monologue and they lend themselves well to the study of how professional readers and speakers respond to the grammatical and rhetorical structure of a text.

Despite the value of the prosodic transcriptions, these corpora also had certain limitations, as Leech (2000) also points out: the *LLC*, although having the richest prosodic annotation, contained conversations that were in practice mostly limited to academic settings – staff and students at London University. The *SEC* contains mainly scripted or prepared public speech, with a few scripted or highly prepared dialogues. More recent compilers have aimed to redress this balance by creating corpora covering a wider range of speaking styles and activities. One example is the *ICE* project, an ambitious project to assemble corpora of different Englishes, the most advanced component of which is the *ICE-GB* (Nelson, Wallis and Aarts 2002). This contains 1 million words of Southern British English, of which 600,000 words are transcribed speech, including informal conversations, broadcast discussions and interviews, debates and cross-examinations, commentaries and speeches. The transcription is orthographic, with a few prosodic features such as hesitations and pauses. An excellent feature of this corpus is that the sound files are not only accessible, but can be aligned with the transcription, so that each line of a concordance, for example, can be listened to in sequence. So while no prosodic annotation of the data is available, anyone interested in studying the text in relation to its phonetic realisation can do so, providing their own prosodic analysis as required (see for example Wichmann 2004).

While the breadth of the *ICE* project is impressive, other subsequent developments aimed for even greater quantities of text, both written and spoken. This generation of ‘mega-corpora’ was often driven by publishers wanting better reference materials. The *Cobuild Corpus* and the *BNC*, for example, contain 20 million and 10 million words of transcribed speech respectively. However, the sound files are not accessible for either of these corpora, except for the subsection of the *BNC* that comprises teenager speech, available separately as the *COLT Corpus*. On the whole, then, despite the increased size and the wide variety of speaking styles etc., the needs of those wishing to study the sounds of speech are being met less well than before.

2.1 The importance of sound

The notion that *how* something is said is as important as *what* is said was central to the work of Gail Jefferson, and to her method of transcribing speech for conversation analysis (see e.g. Jefferson 1985). Her system was driven by the view that in the study of conversational interaction it was crucial to record all audible aspects of the speech and not just the plain text. This included non-speech vocalisations (e.g. laughter), paralinguistic effects (lengthening of syllables, emphatic stress) and segmental features of pronunciation that somehow deviated from an expected ‘norm’. It now seems over-ambitious to hope to subvert orthography for such complex purposes, and it also led to some curious transcriptions, often reminiscent of eye-dialects intended to represent non-standard speech in fiction but which actually record highly predictable forms, such as weak forms, that add nothing to the analysis (*tuh* for *to*, *fuh/fer* for *for*, *ya* for *you*). There is now an encouraging trend in Conversation Analysis (CA)

circles to move away from such a transcription and to make a distinction between the notes taken by the researcher – the working transcription – and the version used for presentation (see Walker 2004). This trend is reinforced by work in edited collections such as those by Couper-Kuhlen and Selting (1996) and Couper-Kuhlen and Ford (2004). These contain work at the interface between phonetics and conversation, and show that there is far more in the speech signal to be accounted for than could ever be captured in an orthographic transcription. The greatest disadvantage of Jefferson's system, or in fact any system of transcription, is that the more detail is included, the greater the danger that the transcription acquires the status of primary data; in fact, the original recordings, while still only a partial representation of the event in that visual information is absent, are primary, and any transcription secondary. Despite shortcomings, however, Jefferson's work was of vital importance in showing that, at least in an interactional setting, prosodic, paralinguistic and non-linguistic features had a role to play in negotiating meaning.

So how can we proceed? First of all, regardless of what corpus compilers feel they can or cannot provide in the way of annotation, we need to make absolutely clear that a spoken corpus is a collection of speech, and that the recordings themselves remain the primary data. No amount of transcription, whether manual or automatic, should mislead the user into thinking the recordings are no longer relevant. It is therefore highly regrettable that a corpus like the *BNC*, obviously of enormous value for many types of linguistic research, was compiled without thought to the accessibility of the sound files. According to Burnard (2002), the procedures involved in anonymising the data have made it difficult if not impossible to approach the participants anew to ask for permission for sound files to be released. The first priority, then, should be to ensure that sound files are available for those who want to use them. Secondly, assuming that they are available, the question is how they should be transcribed. It is again unfortunate that even where recordings are accessible, orthographic transcriptions continue to be created completely separately from the sound, meaning that the alignment of text and sound has to be painstakingly re-created afterwards. This is the case, for example, with the *SEC*, which was later digitised as *MARSEC* (*MACHine-Readable SEC*) (Roach et al. 1994) and the existing transcription embedded in the electronic sound files. In subsequent projects, such as *Aix-MARSEC* (Auran et al. 2004), further annotations have been added, all of which are time-aligned with the sound from the start. Given the wide availability of annotation software, the deliberate breaking of the original link between meaning and sound is unnecessary and short-sighted.

The knowledge that prosody is important for the analysis of meaning in spoken language has generated a growing area of research at the interface between discourse and phonetics, based on the theoretical foundations of CA but employing the technical expertise of trained phoneticians. The current work in the CA framework has already been mentioned, but there is also considerable interest in the speech community in how discourse structures and pragmatic acts are realised phonetically, albeit from a more static and structural perspective. This

requires both discourse/pragmatic and prosodic labelling, but there are considerable problems involved. The biggest drawback is that prosodic annotation especially is labour-intensive, time-consuming, and requires skill and training. Secondly, there is disagreement over transcription systems, and those that exist (the British system of tones and the now most widely used autosegmental system of pitch targets) still do not necessarily capture what some analysts want to study. It is therefore understandable that it has become the norm for the annotation of spoken data to be left to the users.

3. Prosodic annotation

3.1 Auditory

Over the centuries, various methods have been devised to transcribe English prosody. Early systems used staves and notation similar to musical notation. In this century, we have had interlinear pitch notation, ‘tadpole’ notation (again reminiscent of notes in music but without a staff) and ‘crazy type’ (see Williams 1996 for an overview). Most of these systems simply tried to represent the speech melody with no indication of which undulations were significant and which were not. Others, e.g. the tadpole notation, were slightly more reductive in that they indicated stressed and accented syllables, and their pitch direction, while not necessarily indicating intervening pitch patterns. This was and remains a useful shorthand for capturing auditory impressions. The first attempt to represent a prosodic system was that of Crystal (1969) who devised a representation scheme for capturing melody, pause and prosodic phrase boundaries. He also added to his transcriptions impressionistic labels referring to paralinguistic features such as tempo, loudness, voice quality and pitch range.

This was the basis of the system used to transcribe the LLC – a complex annotation that employs a wide range of symbols and is consequently difficult to read. The *SEC* transcription, on the other hand, used a much reduced set of iconic symbols, indicating for the most part only the pitch movement on stressed and accented syllables, and the position of major and minor phrase, or tone-group, boundaries. Both these systems are based on the British model of intonation, in which the smallest unit is a pitch movement (rise, fall, fall-rise, level etc.). They are also purely auditory, given that they were developed before the widespread availability of instrumental analysis and thus the possibility of comparing what we hear with the visible pitch trace or fundamental frequency contour.

This system of annotation has now largely been discarded in favour of that derived from the autosegmental system, in which the smallest unit of analysis is the pitch target, on one of two levels – High or Low. The pitch movements of the British model are in fact interpolations between targets: a fall, for example, is the movement from a High (H) target to a Low (L) target. This is the system employed in most work on prosody in which annotation is used, mainly for applications in laboratory phonology or speech technology. The identification of

pitch targets, or ‘tones’ (not to be confused with the British ‘nuclear tone’) combines with a system of indicating boundaries of different strengths to give the so-called ToBI system of annotation (Tones and Break Indices). This kind of annotation is also auditory, although the visible record (wave form, fundamental frequency trace) is used alongside, to check auditory perceptions. This suggests, of course, that there is a ‘correct’ analysis in which the speech signal is the final arbiter. In fact, although visible changes of pitch direction may help to corroborate or to challenge auditory impressions, only the human analyst can identify a ‘stressed’ or accented syllable, which is then associated with and labelled as a ‘starred’ tone (H* or L*), depending on whether the prominence begins with a high or low target.

Both systems are time-consuming and rely heavily on the skill and training of the transcriber, or in the case of large corpora, the many different transcribers, needed for the work. There are also inevitable disagreements between transcribers, which call into question to some extent the reliability of the transcription. The labelling of data, whether pragmatic, phonological or paralinguistic is no longer contained in separate files, with the consequent loss of alignment with the sound. Instead, software is used that provides multiple annotation tiers linked to the speech signal. These annotation tiers have been used for the annotation of specialised corpora such as the *Reading Emotion Corpus* (Greasely et al. 1995) and the *Oxford/Cambridge IViE Corpus* (Grabe et al. 2001), in each case covering far less speech data than is contained in recent corpora such as the *BNC*. The experience of these transcribers is that the work is so time-consuming and consequently expensive that they simply cannot afford to do more than a small representative sample. The need to employ teams of transcribers, involving intense training periods and frequent consistency checks, can only be accommodated by large-scale, commercially supported projects.

3.2 Automatic

Just as it has been possible to develop programs that automatically tag and parse written text, it is to some extent also possible to capture some features of the speech signal automatically. The developers of the *Aix-MARSEC Corpus* (Auran et al. 2004), for example, used a dictionary look-up to retrieve the phonemic representation of each word in its citation form, and then applied a series of connected speech rules to generate a phonemic transcription of the text. The extent to which this is useful depends of course on the degree of accuracy and the narrowness of the phonetic transcription required. It does, however, allow a reasonable syllabification to be carried out in alignment with the signal.

An alternative method is to apply automatic speech recognition software to the sound files, but this again depends on the degree of accuracy required, and in part, of course, on the quality of the sound files themselves. Naturally-occurring speech can be noisy, making instrumental analysis, and hence the application of ASR, problematic. The automatic extraction of prosodic features is limited: it is possible, for example, to identify periods of silence that can be related to pauses (it would be normal to set a threshold measure of around 200ms

in order to avoid counting segmentally-induced breaks such as the closure phase of a plosive consonant). This information, together with a syllable count derived from automatic syllabification, enables automatic calculation of speech rate or articulation rate.² It is also possible to extract global features of pitch: pitch range, and also pitch maxima and minima. This information can be usefully related to discourse information (speech act, discourse move) and to other annotation categories, if available. Since there is always some manual labelling, or at least manual correction, involved, there is nonetheless a limit to the quantity of data that can be analysed in this way.

While the prosodic features that can be captured automatically are useful in identifying some discourse- or interaction-related prosodic behaviour, especially longer-term effects, these are not easily related to phonological categories. It still requires human labellers to identify stressed or accented syllables, and with the help of visual information extracted from the signal, to label pitch targets as High or Low. Schriberg et al. (1998) compared the automatic identification of prominent syllables with hand labelling and found an accuracy of only 31.7%. Their work is a good example of how labour-intensive the large-scale prosodic analysis of spoken discourse can be: in addition to carrying out automatic extraction of utterance duration, pitch range, pitch movements, energy and speech rates, 10 authors employed a team of five intonation labellers and eight dialogue labellers.

4. So why do we need the sound files, if transcription is so problematic?

Given that prosodic annotation requires such time and effort, we must ask if it is really worthwhile in the context of corpus linguistics. It could be that other methodological approaches are more appropriate for the study of the relationship between prosody and meaning, in particular pragmatic and discursal meaning. Qualitative work in CA, using detailed case-studies, has been able to show that the negotiation of interaction is highly dependent on prosodic and segmental phenomena. Turn-taking mechanisms, for example, involve a variety of vocal signals from pitch to voice quality. Such research has also shown that interpersonal attitudes can be signalled prosodically, as for example in the work on affiliation and disaffiliation by Couper-Kuhlen (1996). More quantitative research on discourse in the speech community, based partly on experimental data and partly on large-scale corpus studies using automatically retrieved data, has shown that there are systematic links between prosody and both the segmentation of discourse and the indication of information status (e.g. given and new), and also between prosody and pragmatic acts (e.g. Schriberg et al. 1998, Venditti and Hirschberg 2003).

It is therefore tempting to assume that both these extremes, detailed case-studies on the one hand, and large-scale quantitative analyses on the other, leave little to be done on the basis of the kind of representative spoken corpora that are used in the corpus linguistics community. I would argue, however, that the rich

information that accompanies them, either in terms of tagging and parsing or in terms of demographic and situational information (e.g. as in *ICE-GB*) provides a unique source of contextual information on which to base studies of human verbal behaviour. This contextual knowledge informs individual studies of *LLC* and *ICE-GB* (e.g. Aijmer 1996, Wichmann 2004), which have shown systematic links between pragmatic acts and prosody. Wichmann (2004), in a study of *please*-requests in *ICE-GB*, found two broad intonation patterns with which such requests are realised, and on the basis of contextual information showed that they are clearly distributed according to the power relations between speaker and hearer. Thus a *please*-request with a low terminal (ending low), e.g. *could I have the next \slide please*, has the force of a command and is used in asymmetrical situations (by a more powerful speaker). A *please*-request with a high terminal, on the other hand, is used in symmetrical situations and has the force of a more tentative request. In an utterance such as *could I have a glass of \water please*, the high terminal underlines the interrogative form and thus allows the addressee the option not to comply. The orthographic transcription of these *please*-requests would fail to show the different force of otherwise equally conventionalised utterances.

Why do we need corpora? Surely, small case-studies or even laboratory experiments would provide similar information? Firstly, laboratory experiments can only be set up around parameters that are already known. The pattern suggested by a corpus study may well be replicated experimentally, but without the prior insights gained from corpus investigation no-one would have thought to test it. Secondly, many pragmatic effects are the result of deviance of some kind – a mismatch between a message and its situational context, for example, generates an inferencing process to identify what is meant that is not said. The *please*-request spoken as a command would be perceived as rude or overbearing if the speaker were not in a position of power over the hearer, while the tentative *please*-request spoken by someone in authority might be perceived as weak, indecisive or ineffectual. In order to identify deviance, or mismatch, we need to identify what constitutes a norm or a ‘match’, and it is for this reason that the quantitative approach afforded by large spoken corpora is invaluable. It is not enough to establish in a few randomly chosen sections of conversation that participants respond similarly to certain phonetic features, segmental or prosodic. With a broader notion of context than is acceptable to CA, e.g. demographic and situational information, it is possible to identify correlations between prosodic patterns and situation (e.g. Speaker-Hearer relationship). On the basis of such information, we can then show that some perceived interpersonal meanings or attitudes arise from inappropriate usage. Corpus data is therefore interesting not only for high-frequency events that one could regard as ‘normal’ behaviour, but also for low frequency occurrences. Wichmann (2005) in a study of emotional requests in *ICE-GB*, revealed that the pattern of *please*-request with the lowest frequency of occurrence was used only in very particular contexts. Thus *please help your\self*, with a high onset on *please* and a falling tone on *self* is a neutral invitation, whereas *\please help your\self*, with a falling tone on *please*, is an

emphatic, emotional exhortation or plea. This is an example of affective or attitudinal meaning that would not be captured by an orthographic transcription alone.

5. Theoretical implications of corpus based prosody research

The corpus-based study of prosody in discourse does not merely have descriptive value, but can contribute in a number of ways to theoretical discussions. In this section I shall discuss first the contribution to the study of language change, then to phonology and finally to psycholinguistics.

5.1 Language change

Researchers who espouse a usage-based functional approach to language (e.g. Bybee 2001), while not rejecting the results of structuralist approaches, believe that such insights need to be complemented by a study of usage, on the assumption that the social and interactional uses to which language is put also play a part in the emergence of structures, both phonological and grammatical. An important aspect of language use is frequency, and Bybee argues that frequency is a significant factor in language change. One of the effects of frequency is habituation, and this can bring with it a change of meaning: words or phrases that once had propositional meaning become ritualised, shifting in use from propositional to interpersonal or discursal meaning. This process is referred to as ‘grammaticalisation’ (see e.g. Traugott and Dasher 2002). One of its most cited features is that of an attenuation of form: commonly cited examples are *going to* > *gonna*, *want to* > *wanna*. Wichmann (2006) suggests that the segmental reductions commonly observed are the result of loss of prosodic prominence. Since prosodic prominence is directly related to information value (see Pierrehumbert and Hirschberg 1990), any process of semantic change that involves the loss of propositional meaning is likely to be accompanied by a concomitant loss of prosodic strength. Words that are not stressed or accented tend to be shorter than the same words given prominence, and the result is less time for careful articulation, leading to incomplete or overlapping gestures.³

A corpus-based example of how this relates to the process of grammaticalisation is a study of the word *sorry* (Aijmer 1996). Aijmer’s study, based on the LLC, shows that the most frequent use of the word is not as an apology but as a marker of conversational difficulty (e.g. *Sorry what did you say? Sorry who?*). In other words, there are contexts in which it is losing its propositional meaning and becoming a discourse marker. Given the tendency for grammaticalisation processes such as this to be accompanied by a reduction in form, it is not surprising that a study (based on the *ICE-GB*) of the prosodic and segmental realisation of *sorry* (Wichmann 2006) showed that when used as a discourse marker, *sorry* was unstressed, short and frequently reduced to [soi], i.e. with loss of *r* colouring. Of course, all tokens in the corpus, regardless of their realisation, are transcribed as *sorry*, since there is no orthographic convention that

would capture the reduced version. Without the availability of the original sound files, this study of the phonology of grammaticalisation would not have been possible.

5.2 Phonology

It can be argued that prosodic transcriptions should be left to the user of spoken corpora, and that is indeed current practice, even where the sound files are available, as with the *ICE* corpora. This, of course, means that only expert prosodists or phoneticians can study the realisation of utterances. The advantage of those corpora already transcribed, such as the *LLC* and *SEC*, lies in the fact that others can also make use of the transcriptions, and, although they have to rely on the decisions of the transcriber, many useful generalisations can be made on this basis. However, there is a further reason why prosodic transcription should be encouraged: it is to test out the validity of the phonological model. Transcriber disagreements mentioned above (section 3.1) may be troublesome to those who aim for consistency as a firm basis for their analyses. However, they are also an important source of phonological information. Herman and Tevis McGory (2002) have observed that many disagreements between transcribers are systematic, and find that this is related to the conceptual similarity of the categories. Their conclusion is that confusability and conceptual similarity raises doubt as to the phonological status of some categories, and that these should be subject to reanalysis.

5.3 Psycholinguistics

Finally, there is a growing body of research into the role played by prosody in the processing of discourse, including on-line interpretation and the relative cognitive loads involved. Venditti and Hirschberg (2003) point out that most of the psycholinguistic information we have so far comes from experimental data, in which experienced speakers and listeners were used, but we have as yet little knowledge of the behaviour of naïve listeners. The role of participants is of course already crucial to work in CA, their observable behaviour providing the only permissible evidence, in this framework, of the function of speaker behaviour. There remains, however, much to be learnt about hearer processing – how prosody aids comprehension and over what domains.⁴ Spoken corpora provide naturally occurring data from which testable hypotheses may be derived.

6. The future

I have argued above that there are several good cases for making sure that the availability of sound files from which spoken corpora derive is given the highest priority, and for supporting large scale prosodic annotation projects. Prosodically annotated data allows non-phoneticians to incorporate into their analyses categories that they would otherwise not feel confident to identify. Such data also

allows continued investigation into the adequacy of the categories themselves, and this is especially important for phonological categories, since prosodic phonology is relatively young compared with other disciplines. The qualitative work being undertaken within the CA framework, while limited in scope, brings incontrovertible evidence that meanings in interaction are generated not only by what people say but by how people say it. Whether one regards interactional categories as the only admissible ones is the function of one's theoretical stance towards language, but any study of meaning related to prosody, both its production and perception, must take into account how it is used in real life.

This is equally true of a further channel of communication – gesture and body language. Detailed qualitative analysis, including that by Conversation Analysts such as Goodwin (1981) has shown that face to face interaction at least is negotiated by means of multiple signalling. Turn-taking is organised by pitch and voice quality but also by gaze: avoidance of eye contact can mean that despite other possible finality signals (syntactic or prosodic) the speaker is not ready to cede the floor. Directing one's gaze at an interlocutor, on the other hand, may indicate a desire to cede the floor fully or only briefly in order to elicit supportive feedback. Gaze patterns have also been found to have discursive relevance. In a study of interaction between airline pilots while flying it was found that talk related to the activity of flying the plane did not include head turn or gaze toward the other pilot, but any talk of a personal rather than professional nature involved head turn and gaze (Neville 2005). In other words, the status of the stretch of discourse – professional or personal – correlated with different body language.

To understand how humans communicate, we therefore need ideally to gather data that includes all channels of communication – verbal, vocal and corporeal. Much work remains to be done beyond existing case studies to gather quantitative and distributional information about how these channels interact. Only in this way can we claim to have identified normative behaviour against which deviance can be measured. The whole area of 'mixed messages', i.e. when the message conveyed by one signal contradicts that conveyed by another, whether it is the source of humour, irony or impoliteness, remains to be explored, and for that we first need to know what is 'normal'.

The technology for recording and storing multimodal data is available, as is software that allows multiple annotations. However, as long as corpus research is driven (and funded) on the basis of just a few aspects of research, such as grammar and lexis, much valuable information could be lost. This paper is therefore a plea to all corpus developers to look beyond their immediate needs and to be a little more visionary in their approach.

Notes

- 1 At least of the type of corpus containing naturally-occurring speech. I do not include here corpora such as the *IViE Corpus*.

- 2 Articulation rate is speaking time, usually measured in syllables per second, excluding pauses.
- 3 This has long been known as a feature of function words in English, such as *for*, *to* and *and*, which are normally realised in reduced form.
- 4 For example, it is well known that topic initiality is reflected in higher pitched onsets than those of non-initial utterances. However, we do not know if this information alone can be used in real time by listeners. On-line processing may well rely more on the local change in F0 from one intonation phrase to the next, than on a memory for utterance onsets (Hirschberg and Nakatani 1996).

References

- Aijmer, K. (1996), *Conversational routines in English*. London: Longman.
- Auran, C., C. Bouzon and D. Hirst (2004), 'The Aix-MARSEC project: an evolutive database of spoken British English', in: B. Bel and I. Marlien (eds.) *Proceedings of speech prosody 2004*, Nara, Japan. Bonn: ISCA. 561-564.
- Burnard, L. (2002), 'The BNC: where did we go wrong?', in: B. Kettemann and G. Marko (eds.) *Teaching and learning by doing corpus analysis: proceedings of the fourth international conference on teaching and language corpora*. Amsterdam: Rodopi. 51-72.
- Bybee, J. (2001), *Phonology and language use*. Cambridge: Cambridge University Press.
- Couper-Kulen, E. (1996), 'The prosody of repetition: on quoting and mimicry', in: E. Couper-Kuhlen and M. Selting (eds.) Cambridge: Cambridge University Press. 366-405.
- Couper-Kuhlen, E. and C. E. Ford (eds.) (2004), *Sound patterns in interaction: cross-linguistic studies from conversation*. Amsterdam/Philadelphia: John Benjamins.
- Couper-Kuhlen, E. and M. Selting (eds.) (1996), *Prosody in conversation. Interactional Studies*. Cambridge: Cambridge University Press.
- Crystal, D. (1969), *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.
- Goodwin, C. (1981), *Conversational organisation: interactions between speaker and hearers*. New York: Academic Press.
- Greasely P., J. Setter, M. Waterman, C. Sherrard, P. Roach, S. Arnfield and D. Horton (1995), 'Representation of prosodic and emotional features in a spoken language database', in: K. Elenius and P. Branderud (eds.) *Proceedings of the 13th ICPHS*, Stockholm, Sweden. Stockholm: KTH and Stockholm University. 242-245.

- Grabe, E., B. Post and F. Nolan (2001), 'The IViE Corpus Department of Linguistics, University of Cambridge' [online]. Available from: <http://www.phon.ox.ac.uk/~esther/ivyweb>.
- Herman, R. and J. Tevis McGory (2002), 'The conceptual similarity of intonational tones and its effects on intertranscriber reliability', *Language and speech*, 45 (1): 1-36.
- Jefferson, G. (1985), 'An exercise in the transcription and analysis of laughter', in: T. A. van Dijk (ed.) *Handbook of discourse analysis*, vol. 3. London: Academic Press. 25-34.
- Knowles G., A. Wichmann and P. Alderson (eds.) (1996), *Working with speech*. London: Longman.
- Knowles, G., B. Williams and L. Taylor (1996), *A corpus of formal British English speech*. London: Longman.
- Leech, G. (2000), 'Grammars of spoken English: new outcomes of corpus-oriented research', *Language learning*, 50 (4): 675-724.
- 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.
- Nevile, M. (2005), 'When airline pilots look at each other: shifts in gaze and signalling the status of talk for managing attention in pilots' work'. Paper presented at the 9th International Pragmatics Conference (IPrA), Riva del Garda, Italy.
- Pierrehumbert, J. and J. Hirschberg (1990), 'The meaning of intonational contours in the interpretation of discourse', in: P. R. Cohen, J. Morgan and M. E. Pollack (eds.) *Intentions in communication*. Cambridge (MA)/London: MIT Press. 271-311.
- Roach, P. J., G. O. Knowles, T. Varadi and S. C. Arnfield (1994), 'MARSEC: a machine-readable spoken English corpus', *Journal of the international phonetics association*, 23 (2): 47-54.
- Schriberg, E., R. Bates, A. Stoltcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer and C. van Ess-Dykema (1998), 'Can prosody aid the automatic identification of dialog acts in conversational speech?', *Language and speech*, 41 (3/4): 443-492.
- Stenström, A.-B. (1994), *An introduction to spoken interaction*. London/New York: Longman.
- Svartvik, J. (ed.) (1990), *The London Lund corpus of spoken English: description and research*. Lund Studies in English 82. Lund: Lund University Press.
- Traugott, E. C. and R. B. Dasher (2002), *Regularity in semantic change*. Cambridge: Cambridge University Press.
- Venditti, J. J. and J. Hirschberg (2003), 'Intonation and discourse processing', in: M.-J. Solé, D. Recasens and J. Romero (eds.) *Proceedings of the 15th ICPPhS*, Barcelona, Spain. 107-114.
- Walker, G. (2004), 'The phonetic design of turn endings, beginnings and continuations in conversation'. Unpublished PhD dissertation, York.

- Williams, B. (1996), 'The formulation of an intonation transcription system for British English', in: G. Knowles, A. Wichmann and P. Alderson (eds.) *Working with speech*. London: Longman. 38-57.
- Wichmann, A. (2000), *Intonation in text and discourse*. London: Longman.
- Wichmann, A. (2004), 'The intonation of *please*-requests: a corpus-based study', *Journal of pragmatics*, 36 (9): 1521-1549.
- Wichmann, A. (2005) '*Please* – from courtesy to appeal: the role of intonation in the expression of attitudinal meaning', *English language and linguistics*, 9 (2): 229-253.
- Wichmann, A. (2006), 'Prosody and discourse: a diachronic approach'. To appear in *Proceedings IDP05 interface discours-prosodie (discourse-prosody interface)* 8-9 Sept 2005, Aix-en-Provence, France.

Descriptive studies in English syntax and semantics

This page intentionally left blank

An example of frequent English phraseology: distributions, structures and functions

Michael Stubbs

University of Trier

Abstract

One area of linguistics which has developed very rapidly in the last 25 years is phraseology. Corpus study has shown that routine phraseology is pervasive in language use, and various models of recurrent word-combinations have been proposed. This paper discusses aspects of frequent phraseology in English: the distribution of recurrent multi-word sequences in different text-types and the structure, lexis and function of some frequent multi-word sequences. Most of the data come from a major interactive data-base which provides extensive quantitative information on recurrent phraseology in the British National Corpus (BNC). This data-base, available at <http://pie.usna.edu>, has been developed by William Fletcher. Quantitative phraseological data have important implications for linguistic theory, because they show how findings from phraseology can be related to independent findings from other areas of linguistics, including recent studies of grammar and of semantic change. However, the very large amount of data itself poses methodological and interpretative puzzles.¹

1. Section 1

This paper consists of two main sections. Section 1 describes resources for studying frequent phraseology across large corpora, including software for identifying recurrent multi-word sequences and a phraseology data-base. Section 2 discusses a small set of illustrative findings from the data-base along with some of their implications, and points out a few major problems for analysis.

1.1 Phraseology

One main discovery of corpus work over the last 25 years is a pervasive syntagmatic phrasal organisation in language use, which had been largely ignored, first, because it did not fit into either lexis or grammar, and second, because it involved facts about frequency, which were either out of fashion or which could be studied only with computational help. The importance of phraseology had been recognised from the 1930s onwards by a few individuals (such as H. Palmer, A. S. Hornby and J. R. Firth in the UK, and somewhat later by D. Bolinger in the USA), but had been seen by most linguists only as an unrelated collection of oddities. In addition, simple but crucial quantitative methods of studying routine phraseology had been proposed as early as the 1970s (Allén 1975), but had received less attention than they deserved. A crucial change of perspective came in the 1980s and 1990s, when phraseology was recognised as

a pervasive and possibly unified phenomenon. This change is discussed in major reviews by Cowie (1998), Moon (1998), Pawley (2001), Wray (2002) and others. It gained support in approaches to syntax such as pattern grammar (Francis et al. 1996, 1998; Hunston and Francis 2000) and construction grammar (e.g. Goldberg 1995; Michaelis and Lambrecht 1996; Kay and Fillmore 1999; Croft 2001). Pattern grammar and construction grammar are in many ways comparable and compatible, although only pattern grammar has grown out of work on corpus-based phraseology, and there is little if any contact between these two approaches. Many case studies provide detailed analyses of phrasal complexes (e.g. Sinclair 1991; Louw 1993; Channell 2000; Hunston 2002; Stubbs 2001), and Sinclair (1998) proposes a model of extended lexical units. More recent work by Fletcher (2003/2004), discussed below, now gives easy access to quantitative data on phraseology across large corpora, and means that the extent of routine language use and its implications for linguistic theory can be empirically studied. Despite their diverse origins, this work has all concluded that phraseology can shed light on a fundamental linguistic problem: establishing the units of the language, especially units of meaning (Sinclair 2004: 24-48, 131-148).

Evidence that many aspects of phraseology are habitual, unnoticed and often deeply unconscious is that the collocations involved are often not recorded in dictionaries (which have never developed consistent ways of handling syntagmatic patterns of collocation), and have largely been ignored by grammars (which have underestimated the links between lexis and grammar). Speakers often become consciously aware of these patterns, only when they come across a slightly odd collocation.

1.2 Terms and definitions

In order to study phraseology quantitatively, we need to define 'multi-word sequence' in ways which are suitable for corpus work (that is, so that software can find these sequences in texts), and we need a data-base formed from the application of the software to a large corpus.

Out of the many possible concepts of 'multi-word sequence', I will use just three related definitions. First, I will use the term 'n-gram' to mean a recurrent string of uninterrupted word-forms. Software to extract n-grams proceeds through a text or corpus, with a moving window, a given number of words at a time. It keeps a record of how often each n-gram occurs, and orders them alphabetically or by frequency. There are no standard terms for n-grams, which are called 'statistical phrases' (Strzalkowski 1998: xiv), 'recurrent word-combinations' (Altenberg 1998), 'lexical bundles' (Biber et al. 1999), 'clusters' (Scott 1997), 'chains' (Stubbs and Barth 2003) and 'multi-word sequences' (Butler 2003).

We need a second concept which is more flexible than an n-gram. A 'phrase-frame' ('p-frame') is an n-gram with one variable slot. For example, we can study 5-frames, such as *plays a *part in* with its variant 5-gram realisations. The following are extracted from a data-base, described below, constructed from

the *British National Corpus* (*BNC*). The adjectives are listed in descending frequency.

*plays a * part in <large, significant, big, major, vital, essential, key, central, full, great, prominent,...>*

These adjectives are all rough synonyms (in this p-frame), and we can use such data to study the extent to which such frames contain only restricted vocabulary. The two-word ‘collocational frameworks’ identified by Renouf and Sinclair (1991) are one special case of phrase-frames: ‘discontinuous pairings’ which enclose characteristic sets of words. One of their examples is the 3-frame *a * of*, which has frequent realisations such as

*a * of < number, lot, couple, series, variety, group, range, set, pair, list,...>*

I will use a third term ‘PoS-gram’ to mean a string of part of speech categories. This can be defined only with reference to a particular set of PoS-tags which have been used on a particular corpus. I will be talking about PoS-grams extracted from the *BNC*, where one of the most frequent 5-PoS-grams is

PRP AT0 NN1 PRF AT0
= preposition + determiner + singular noun + *of* + determiner
e.g. *at the end of the; as a result of the; in the middle of the*

1.3 ‘Pie’: the *BNC* phraseology data-base

These three terms ‘n-gram’, ‘phrase-frame’ and ‘PoS-gram’ are used by Fletcher (2003/2004) in connection with a large and powerful interactive data-base which he has constructed from the 100-million-word *BNC*. This data-base can be searched in many different ways for quantitative information on recurrent multi-word sequences. Fletcher has extracted from the *BNC* all n-grams, p-frames, and PoS-grams of length 1 to 6, down to a cut-off frequency of 3 for n-grams. (1-grams are individual word-forms, so the data-base can also provide customised word frequency lists of many kinds.) Clicking on a p-frame or a PoS-gram produces a list of its n-gram variants. Clicking on an n-gram produces up to 50 examples of its use in context in the *BNC*. The web-site gives full details of exactly how words are defined, in what ways the data have been normalised, etc. The data-base is available at <http://pie.usna.edu> (‘pie’ = Phrases in English).

The data-base allows different search parameters to be set: for example, the minimum or maximum frequency of multi-word sequences to be retrieved, or whether results are shown with PoS-tags (as defined by the *BNC* coding). Searches can include wildcards (* = any word, ? = any single character). Filters can be set, in order to specify individual word-forms or PoS-tags or both (for the same position in a sequence), or a mixture of both (for the whole sequence). The number of possible combinations here is clearly astronomically high, but a few

simple examples give a rough idea of the range available. It is possible to search for patterns such as the following, in order to generate tailor-made frequency lists of many kinds:

- The most frequent realisations of the verb lemma KNOW: *kn?w** + verb. The pattern **kn?w**, with no PoS-category specified, would give in addition *unknown, knowledge*, etc.
- The most frequent exponents of any PoS-category, e.g. the most frequent lexical verbs, nouns or prepositions (a 1-gram is an unlemmatised word-form). In this way, customised word-frequency lists can be generated.
- 3-grams consisting of adjective + noun + noun. This finds many well known phrases (e.g. *high interest rates, high blood pressure*), including many names of institutions (e.g. *National Health Service, Local Education Authorities, Royal Air Force*).
- 4-frames which do or do not begin with a preposition. Searches can thus include or exclude patterns.
- 5-grams consisting of *PLAY* (lemma) + *a/an* + adjective + *part* + *in*.
- The most frequent PoS-grams of length 5 (as illustrated above).
- Searches can be defined for fuzzy matches, e.g. *~ at the ~ top of* gives *right at the very top of* or *at the very top of* or *right at the top of* or *at the top of*.

Future developments planned by Fletcher (post-2005) include: extending the length of extracted recurrent strings from 6 to 8 orthographic words; providing searches by regular expressions (only wildcards are currently supported); providing data on both frequency and range (defined by a dispersion measure which counts how many text sectors of arbitrary length a multi-word sequence occurs in, as in Leech et al.'s 2001 *BNC* word-frequency lists); providing data on frequency in different text-types (such as spoken – written, fiction – non-fiction, academic – non-academic, etc., via the *BNC* domain categories and via Lee's 2002 revised *BNC* categories); providing lexical association metrics; and including other corpora in the data-base (such as the *Michigan Corpus of Academic Spoken English* and, when it becomes available, the *American National Corpus*). In summary, the data-base is a massive virtual machine for re-arranging data, in order that we can see previously invisible patterns of phraseology.

One simple, but severe, interpretative problem arises from the very large amounts of data involved. For example: over 225,000 5-grams occur five times or more in the *BNC*, and over 1,100 occur a hundred times or more; over 7,800 5-frames occur five times or more, and over 100 occur a hundred times or more. Figures for 3- and 4-grams and for 3- and 4-frames are correspondingly very much higher. This provides a severe problem for description, since it is difficult to know what level of delicacy is appropriate in making generalisations across so much data. The only realistic strategy is to start small: to use a restricted sample to generate hypotheses which can be tested on larger samples.

1.4 Sample questions

So far I have discussed: three definitions of ‘multi-word sequence’ which are very simple, but useful for corpus work; software to identify these multi-word sequences and their frequencies across texts; and corpora; and a very powerful interactive data-base constructed by applying the software to the *BNC*. I have also started to indicate a few aspects of phraseology which can be studied with these very rich resources, but the data-base can be used for many kinds of study, and it will take a long time before we can properly appreciate the full range of generalisations about phraseology which it allows us to investigate. The rest of the paper is therefore an initial brief illustration of some topics for research.

2. Section 2

In any area of study, findings are significant only if they can be linked to other findings. It is best if findings can be causally linked, or if they can be shown to be the same findings seen from a slightly different point of view. In this section, I will give some initial illustrations of these points.

2.1 Text-types

One almost purely descriptive application of n-gram software is to show that different n-grams occur with different frequencies in different text-types. Biber et al. (1999) do this when they compare n-grams (which they call ‘lexical bundles’) in the broad text-types ‘conversation’ and ‘academic prose’. (Other studies are by Milton 1998 and Aarts and Granger 1998). The main idea here is very simple. For example, the frequency of pronouns distinguishes text-types, such as fiction and academic articles, but if we look at the n-grams in which pronouns occur, then the differences between the text-types are much more striking. It is intuitively obvious which of the following 4-grams come from ‘fiction’ and which from ‘academic prose’:

- (1) I don’t want to; I want you to; I don’t know what
- (2) I have already mentioned; I shall show that; I will conclude with

These examples are from Stubbs and Barth (2003), who show that the rank orders and frequencies of these 4-grams are also very different in different text-types. Set (1) are from the 25 top 4-grams in a corpus of ‘fiction’ and all occur almost 30 times per million; the 50 top 4-grams in a corpus of ‘academic prose’ contain no pronouns at all, and set (2) are from much further down the list from ‘academic prose’ and occur 6 times or fewer per million. Similarly, the kind of prepositional phrases discussed below are much more frequent in written academic texts than in spoken language.

2.2 Frequent multi-word sequences in English

Other implications of recurrent and frequent multi-word sequences can be illustrated from a small set of facts which are both unexpected and inevitable (Hardy 1940/1967: 113): unexpected in that native speakers cannot produce the facts from introspection, but inevitable once it is realised why a particular search method finds these phrases. One problem, as noted above, is simply the very large amount of quantitative data which the data-base makes available. We just have to start somewhere by selecting a sub-set of data: one simple starting place is the most frequent multi-word sequences of a given length in the whole *BNC*. For example, the four top 5-PoS-grams in the whole *BNC* (down to a cut-off of 3 for n-grams) are parts of nominal and prepositional phrases, which express spatial, chronological and logical relations:

PRP AT0 NN1 PRF AT0	ca 990 per million (e.g. <i>at the end of the</i>)
AT0 NN1 PRF AT0 NN1	ca 570 per million (e.g. <i>the end of the year</i>)
AT0 AJ0 NN1 PRF AT0	ca 260 per million (e.g. <i>the other side of the</i>)
PRP AT0 AJ0 NN1 PRF	ca 200 per million (e.g. <i>on the other side of</i>)

The following all occur in the twelve top 5-frames in the *BNC* (numbers to the left indicate rank order):

1.	in	the	*	of	the
2.	at	the	*	of	the
3.	to	the	*	of	the
6.	on	the	*	of	the
9.	for	the	*	of	the
10.	by	the	*	of	the
12.	in	the	*	of	a

The other 5-frames in the top 25 are almost all variants of these multi-word sequences (e.g. *of the * of the*; ** the end of the*; *the end of the **; and *the * of the*; *at * end of the*).

I give only approximate average figures above, since the *BNC* consists of 90 million words of written data and only 10 million words of spoken data, and, as noted, different text-types have significantly different phraseology. (In spoken data 5-frames with high frequency verbs are frequent.) Nevertheless, these prepositional phrases are at the top in both written and spoken samples, so this is a good pattern to start with. To make sampling even simpler, and to make sure that we have a well defined and replicable sample, we can start with just the 150 top 5-grams in the whole *BNC* which have the PoS structure

preposition + determiner + singular noun + *of* + determiner
 = adapted *BNC* coding: PRP AT0 NN? PRF AT0

(The data-base supports patterns with wild-cards: NN? = NN0 | NN1 | NN2.) These 5-grams are listed in the Appendix. The most frequent, *at the end of the*, occurs over 45 times per million running words. Frequencies then decrease rapidly. The next most frequent (with occurrences per million words) are: *by the end of the* (19), *as a result of the* (16), *in the middle of the* (15), and *at the time of the* (11). In four reference corpora of written texts of one million words each (*LOB*, *Brown*, *FLOB* and *Frown*), the frequencies of these five top 5-grams are, respectively: *at the end of the* (41, 38, 37, 24), *by the end of the* (10, 6, 21, 11), *as a result of the* (15, 11, 16, 6), *in the middle of the* (15, 15, 22, 19), *at the time of the* (10, 9, 16, 7).

All 5-grams in the Appendix occur at least once per million running words. These 150 types are realised by around 45,000 tokens (in the whole *BNC*). That is, one of these 150 items occurs more than once every 2,000 running words on average. There are, in fact, many more items with the same or similar structures which do not occur on the list: realisations which occur fewer than 100 times each, discontinuous variants (e.g. *on the other side of*), and 4-grams (e.g. *at the end of*) which occur much more frequently than these 5-grams. This is clearly only a tiny tip of the iceberg of English phraseology, but since these multi-word sequences are the most frequent 5-grams, as selected by very simple criteria, we can use them to make hypotheses which can be tested on larger data-sets, and then check whether generalisations here can be convincingly related to other facts.

2.3 Semantics

The nouns in the list in the Appendix are selected from restricted semantic fields. By far the most frequent noun is *end*: in over 10 per cent of the types and over 20 per cent of the tokens. As in all other areas of language use, the list shows a very uneven distribution. The prepositional constructions occur in a cluster at the top of frequency lists of 5-grams, 5-frames and 5-PoS-grams. And these 150 top 5-grams follow a Zipf-type rank-frequency curve (e.g. Zipf 1945). The two top 5-grams (*at/by the end of the*) constitute 13 per cent of all the tokens (in this top set of 150). The ten top 5-grams; with the nouns *end* (x 2), *result*, *middle*, *time*, *top*, *beginning*, *case*, *part*, *form*, constitute over 30 per cent of all the tokens (in this top set of 150).

Some lexical characteristics of these frequent 5-grams are clear: they contain high frequency nouns (almost all singular) from the core vocabulary, especially place, time and logical terms, (and, in spoken data, not illustrated here, a few high frequency verbs). The list consists overwhelmingly of (the beginnings of) expressions which denote wholes and parts of things, especially the centre or the periphery of places and periods of time:

for the duration of the; for the rest of the; in this part of the; for the remainder of the; for the whole of the; since the beginning of the; at the edge of the; etc.

A second set of expressions denote logical or causal connections:

in the case of the; in the event of a; as a result of the; on the basis of the; as a consequence of the; with the exception of the; etc.

A third set, not entirely distinguishable from causal meanings, and rather further down the frequency list, express intentions and/or relations of power or influence, especially between people:

for the benefit of the; under the auspices of the; in the interests of the; for the purposes of the; at the request of the; for the sake of the; under the control of the; at the expense of the; at / in the hands of the; etc.

Another striking feature of the 5-grams is that many are not semantically transparent. Some are, of course, because some prepositional phrases are simply literal place expressions:

<i>in the corner of the</i>	<room, field, ...>
<i>in the direction of the</i>	<river, town, ...>
<i>at the top of the</i>	<stairs, hill, ...>

Many of the nouns are also body terms, but they can only very rarely be interpreted as body terms in this grammatical construction. This is well known from work on diachronic semantic shifts: see below.

<i>in the heart of the</i>	<city, forest, ...>
<i>at the back of the</i>	<house, book, ...>
<i>by the side of the</i>	<road, bed, ...>
also: <i>face, foot, hands, head</i>	

For many other 5-grams, although the etymology is transparent, no literal interpretation is possible, since the noun is delexicalised, and the meaning of the resultant n-gram cannot be derived purely compositionally. The last attested example below shows just how delexicalised such nouns can be: the writer was apparently not aware of any logical contradiction between the nouns.

<i>at the heart of the</i>	<matter, problem, ...>
<i>on the eve of the</i>	<battle, election, ...>
<i>in the eyes of the</i>	<law, public, ...>
<i>in the wake of the</i>	<riots, scandal, ...>
<i>at the height of the depression</i>	

2.4 Pragmatics

Several of the expressions have pragmatic connotations. The conventionally negative evaluative meaning of *at the hands of the* is clear in examples such as:

suffered humiliation *at the hands of the* Puritans
 experienced persecution *at the hands of the* regime

In other cases the connotations are less obvious to introspection. For example, the 4/5-gram *in the middle of (the)* often (but not always) occurs when the speaker is complaining about something (usually someone else's behaviour) which is unexpected and/or inappropriate, and which has happened where it normally doesn't and/or shouldn't. One hint of this frequent evaluative connotation is that in the whole *BNC* the 4-gram *in the middle of* is most frequently followed by *the night* (n = 277) and *the road* (n = 85). The following are illustrative examples:

he gets called out right *in the middle of* the night
 they just left it *in the middle of* the road
 lying strewn on the floor, *in the middle of* the room for everyone to see!
 I'll give you a ring back. ... we're *in the middle of* eating
 they live in a ghastly little bungalow *in the middle of* nowhere

A corpus can tell us in this way which sequences are frequent, but an explanation of why they are frequent can come only from texts. It is not surprising that expressions for place, time, cause and intention are amongst the most frequent in the language, because these are precisely the relations which we need in order to reconstruct plausible sequences of events, and therefore to make sense of connected discourse. The large majority of the 5-grams in the Appendix end with a definite article, signal an anaphoric referent, and therefore contribute to text cohesion. Several exceptions which end in an indefinite article refer to a hypothetical future (e.g. *in the event of a*).

These observations corroborate generalisations in other studies about the functions of recurrent multi-word sequences. For English data (spoken, from the *London-Lund Corpus*), Altenberg (1998) makes several generalisations about the 'pragmatic specialisation' of recurrent word-combinations. He identifies frequent syntactic constructions (including nominal and prepositional groups), and shows that many routine expressions have the 'conventionalised discourse function' of presenting information in textual frames. For English and Spanish data (spoken and written), Butler (1998, 2003) makes similar observations. He also notes that many frequent multi-word units are nominal or prepositional phrases, that rather few of these phrases encode representational meanings, except for expressions of time and place, and that many frequent sequences express speaker-orientation and information management.

We have to remember, however, that different observational methods lead to different findings: if we use a microscope we will discover small things, but if we use a telescope we will discover distant things, and if we use x-rays we will discover what is inside things. The software picks out multi-word sequences which are both frequent and widely distributed, and which are therefore, by definition, not tied to the topic of individual texts. They are used by speakers, irrespective of what they are talking about, in order to organise their discourse, and therefore contain many markers of point of view, topicalisation, and the like. So, as well as seeing these generalisations as an empirical finding which is induced from the data, we can also look at things the other way round. The criteria of high frequency and wide range automatically capture predominantly non-representational phraseology. It is not that the findings are a mere artefact of the method. In retrieval tasks, success is measured in terms of precision and recall. If we are searching for such phraseology, then this method has fairly high precision (much of what is found is relevant). Speakers constantly refer to times and places, in routine ways, in order to organise both narrative and non-narrative texts. These prepositional phrases are a recurrent way of organising information. We do not know how high the recall is (whether the method finds most of what is relevant), but recall is always much more difficult to check, since we cannot observe what is not found.

Summarising some of these points: what I have described is evidence of a construction which has a well-defined syntax and lexical-semantic characteristics. It has prototypical (high frequency) exemplars. It contains vocabulary from restricted lexical classes. It has specialised pragmatic functions, primarily in managing information and structuring text. It is an idiomatic form-meaning complex, in the sense of construction grammar, although the construction is rather less specific than several examples which have been discussed in the literature (Fillmore et al. 1988; Goldberg 1995; Michaelis and Lambrecht 1996; Kay and Fillmore 1999; Croft 2001 et al.). I have however glossed over several problems in correctly identifying the boundaries of the underlying construction: see below.

2.5 Relations to work on language change

The vocabulary of these recurrent 5-grams has much in common with vocabulary which has long been identified as particularly prone to certain kinds of semantic shift. The prepositional phrases which I have discussed are very frequent and they frequently contain nouns from certain semantic fields. These two factors make these phrases a plausible context for semantic change, and the nouns are indeed frequently delexicalised. There is extensive evidence from many languages that words in certain lexical classes undergo predictable semantic shifts. For example, body part nouns often become place nouns (e.g. *back*, *side*), and these nouns often shift further, to become place adverbials and discourse markers (e.g. *beside*, *besides*). (Rissanen 2004 discusses the example of *side* in detail.) Another example is the word *face*, which is used both 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.

These are examples of well documented uni-directional diachronic processes which affect predictable classes of words in the basic vocabulary, and which involve shifts from concrete to increasingly abstract expressions, with a weakening of semantic meaning, and a corresponding strengthening of pragmatic meaning (e.g. speaker attitude). Typical developmental tracks have been identified and labelled in slightly different but related ways (e.g. by Traugott and Heine 1991; Hopper and Traugott 1993).

concrete / literal > abstract / metaphorical
 body part > locational term > temporal term > discourse term
 locational > temporal > logical > illocutionary
 propositional / extralinguistic > expressive / attitudinal

It is always a hint that we are on the right track if it can be shown that two apparently distinct sets of facts are different ways of looking at the same thing. Quantitative phraseological data can now provide many examples to support the hypothesis (proposed by Hopper and Traugott 1993) that predictable kinds of semantic shifts take place in local grammatical constructions. This is true especially of the semantic weakening of nouns in frequent prepositional constructions, and the corresponding strengthening of pragmatic meanings (to form text-structuring expressions and/or conventional evaluative connotations). The papers in Lindquist and Mair (2004) discuss in detail the implications of corpus data for studies of grammaticalisation.

2.6 Conclusions and implications

I will conclude by summarising a few findings which seem well established, but also by pointing to major unsolved analytic problems. Some of the following points are well known, but if discoveries about phraseology are important, then it is due to their consequences for a wide field of language study, so it is useful to try and state explicitly some theoretical implications.

Corpus studies have discovered large numbers of new facts, regularities (where people had previously seen only irregularities), and relations between things (which had previously seemed independent). The broadest significance of such findings is for a theory of idiomatic language. Pawley and Syder (1983: 191-193) pointed out that “native speakers do not use the creative potential of syntactic rules to anything like their full extent”, but instead have standard ways of talking about culturally recognised concepts. This observation can now be studied in empirical quantitative detail.

There are more specific implications for how we model the vocabulary, grammatical units and textual structure. The multi-word sequences which I have identified consist of predictable classes of words which are used in predictable ways, they are similar to units which have been discovered independently in other

areas (such as construction grammar), and they have predictable functions (such as managing information).

- (i) If we are thinking of words: many words are frequent because they occur in frequent phrases (Summers 1996: 262-263; Sinclair 1999: 162). This can be measured as their constructional tendency (Allén 1975). The concept of “word frequency” therefore needs reinterpretation.
- (ii) If we are thinking of phrases: many phrases are frequent because they are conventional ways of expressing common meanings (Pawley and Syder 1983). Studying these phrases alters our understanding of what speakers have to infer versus what is conventional and just has to be known.
- (iii) If we are thinking of texts: many frequent phrases express the relations of place, time, cause and intention which are essential for understanding connected discourse (Altenberg 1998; Butler 2003).
- (iv) If we are thinking of the vocabulary: Bloomfield’s (1933: 274) famous description of the lexicon as “a list of basic irregularities” is too pessimistic. (So is the notion of phraseology as just a list of oddities about fixed phrases and idioms.) Many generalisations about vocabulary have now been discovered.
- (v) If we are thinking of language change: quantitative phraseological data can help explain why words which occur frequently in well-defined grammatical constructions undergo predictable semantic shifts.

The quantitative phraseological data which I have illustrated have been available for only a very brief period of time. They provide evidence for related questions of linguistic description and theory, but their full implications will become clear only after concentrated study by many people, and lurking in the background of this paper is the deep problem which I have alluded to only in passing: the appropriate criteria for identifying linguistic units.

A first version of the problem is formulated as point (i) above: there is something slightly wrong, logically, with the concept of ‘word frequency’. Many words are frequent because they occur in frequent phrases, not only in units which are tagged as ‘multi-words’ in the *BNC*, but also in highly frequent n-grams which are (presumably) not multi-words. However, the problem must also be formulated from the other end: the linguistic status of frequent multi-word sequences is not at all clear, since frequent recurrence is only one criterion for a linguistic unit.

In this paper I have discussed only a small and rather arbitrarily selected sample of n-grams. I could have selected longer n-grams and, indeed, in some cases it is clear that the 5-grams which I have discussed are part of longer semantic units. For example, there are around 4570 occurrences of *at the end of*

the in the *BNC*, of which around 17 per cent are in the phrase *at the end of the day* (the percentage is higher in spoken data), and most of these are used in the sense ‘when everything has been taken into account’. Further occurrences are in the phrase *light at the end of the tunnel*, mostly in the sense that ‘the solution to a problem is in sight after long-term difficulties’. Similarly, many occurrences (but not all) of *in the wake of the* are followed by nouns with clearly negative connotations, such as *crisis, controversies, riots, row, scandal, war*. This is semantic evidence (in the form of non-compositionality and/or of a semantic prosody), that the 5-gram is part of a longer linguistic unit.

Conversely, it might seem more plausible to take recurrent 4-grams (e.g. *at the front of, in the wake of*) as the units to investigate, since they look like complex prepositions (e.g. analogous to *in front of, following*). The coding of ‘multi-words’ in the *BNC* is not transparent in this area: *in the light of* and *on the part of* are the only two strings with this structure (preposition + determiner + noun + *of*) which are coded as multi-words, although other 4-grams (e.g. *in the case of, in the face of*) are also frequent and semantically non-compositional. But then reference grammars do not agree as to whether such strings are complex prepositions (see the lack of consensus on this issue across Halliday 1994: 212; Quirk et al. 1985: 670-673; Biber et al. 1999: 75-76, 113; and Huddleston and Pullum 2002: 617-623). All of these grammars usefully discuss different criteria for distinguishing complex prepositions from freely constructed sequences, including syntactic and semantic indivisibility, restricted variability, restricted modification, fossilisation, lack of semantic transparency, and all, to different extents, admit that the distinction is a matter of degree. However, Huddleston and Pullum (2002: 617-623) discuss the “remarkable profusion of idiomatic and semi-idiomatic constructions into which prepositions enter”, and then reject the “complex preposition” analysis entirely, on the grounds that its “initial intuitive appeal” is based on unreliable semantic criteria (Huddleston and Pullum 2002: 621). High frequency of recurrence is only one type of evidence in favour of the linguistic status of a unit, but it is striking that none of these reference grammars use frequency data at all.

Quantitative data on recurrent linear strings of word-forms raise the classic problem of how to identify linguistic units. There is still a wide gap in this area between evidence and interpretation, and between the demand for systematic empirical description (where corpus studies are particularly strong) and theory building (which will take time to catch up with the vast amounts of new observational data which have become available).

Note

- 1 This paper reports work which was done in collaboration with Isabel Barth (Stubbs and Barth 2003) and Katrin Ungeheuer. The software I have used in studying n-grams and p-frames was written by Isabel Barth and Bill Fletcher. I am especially grateful to Bill Fletcher, for much discussion, for

the use of a beta-version of his n-gram and p-frame software, and for access to early versions of his *BNC* data-base (Fletcher 2003/2004). For comments on previous drafts I am also grateful to Chris Butler, Joanna Channell, Naomi Hallan and members of the BAAL Special Interest Group on corpus linguistics, to whom an earlier version of the talk was given at Birmingham University in April 2004.

References

- Aarts, J. and S. Granger (1998), 'Tag sequences in learner corpora', in: S. Granger (ed.) *Learner English on computer*. London: Longman. 132-141.
- Allén, S. (1975), *Nusvensk frekvensordbok*. Stockholm: Almqvist and Wiksell.
- 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., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman grammar of spoken and written English*. London: Longman.
- Bloomfield, L. (1933), *Language*. London: Allen and Unwin.
- Butler, C. (1998), 'Collocational frameworks in Spanish', *International journal of corpus linguistics*, 3 (1): 1-32.
- Butler, C. (2003), 'Multi-word sequences and their relevance for models of functional grammar', *Functions of language*, 10 (2): 179-208.
- Channell, J. (2000), 'Corpus-based analysis of evaluative lexis', in: S. Hunston and G. Thompson (eds.) *Evaluation in text*. Oxford: Oxford University Press. 38-55.
- Cowie, A. P. (ed.) (1998), *Phraseology*. Oxford: Oxford University Press.
- Croft, W. (2001), *Radical construction grammar*. Oxford: Oxford University Press.
- Fillmore, C., P. Kay and M. C. O'Connor (1988), 'Regularity and idiomaticity in grammatical constructions', *Language*, 64: 501-538.
- Fletcher, W. (2003/2004), *Phrases in English* [online]. Available from: <http://pie.usna.edu>.
- Francis, G., S. Hunston and E. Manning (1996, 1998), *Grammar patterns*, volumes 1 and 2. London: HarperCollins.
- Goldberg, A. (1995), *Constructions*. Chicago: Chicago University Press.
- Halliday, M. A. K. (1994), *An introduction to functional grammar*. London: Arnold.
- Hardy, G. H. (1940/1967), *A mathematician's apology*. Cambridge: Cambridge University Press.
- Hopper, P. and Traugott E. (1993), *Grammaticalization*. Cambridge: Cambridge University Press.
- Huddleston, R. and G. K. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

- Hunston, S. (2002), *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hunston, S. and G. Francis (2000), *Pattern grammar*. Amsterdam: John Benjamins.
- Kay, P. and C. Fillmore (1999), 'Grammatical constructions and linguistic generalizations: the *What's X doing Y?* construction', *Language*, 75 (1): 1-33.
- Lee, D. Y. W. (2002), 'Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the *BNC* jungle', in: B. Kettermann and G. Marks (eds.) *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi. 247-292.
- Leech, G., P. Rayson and A. Wilson (2001), *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.
- Lindquist, H. and C. Mair (eds.) (2004), *Corpus approaches to grammaticalization in English*. 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*. Amsterdam: John Benjamins. 157-176.
- Michaelis, L. and K. Lambrecht (1996), 'Towards a construction-based theory of language functions', *Language*, 72: 215-247.
- Milton, J. (1998), 'Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment', in: S. Granger (ed.) *Learner English on computer*. London: Longman. 186-198.
- Moon, R. (1998), *Fixed expressions and idioms in English: a corpus-based approach*. Oxford: Clarendon.
- Pawley, A. (2001), 'Phraseology, linguistics and the dictionary', *International journal of lexicography*, 14 (2): 122-134.
- Pawley, A. and F. H. Syder (1983), 'Two puzzles for linguistic theory', in: J. C. Richards and R. W. Schmidt (eds.) *Language and communication*. London: Longman. 191-226.
- Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985), *A Comprehensive grammar of the English language*. London: Longman.
- Renouf, A. and J. Sinclair (1991), 'Collocational frameworks in English', in: K. Aijmer and B. Altenberg (eds.) *English corpus linguistics*. London: Longman. 128-143.
- Rissanen, M. (2004), 'Grammaticalization from side to side: on the development of *beside(s)*', in: H. Lindquist and C. Mair (eds.) *Corpus approaches to grammaticalization in English*. Amsterdam: John Benjamins. 151-170.
- Scott, M. (1997), *WordSmith Tools manual*. Oxford: Oxford University Press.
- Sinclair, J. (1991), *Corpus concordance collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1998), 'The lexical item', in: E. Weigand (ed.) *Contrastive lexical semantics*. Amsterdam: John Benjamins. 1-24.
- Sinclair, J. (1999), 'A way with common words', in: H. Hasselgard and S. Oksefjell (eds.) *Out of corpora*. Amsterdam: Rodopi. 157-179.

- Sinclair, J. (2004), *Trust the text*. London: Routledge.
- Strzalkowski, T. (ed.) (1998), *Natural language information retrieval*. Dordrecht: Kluwer.
- Stubbs, M. (2001), *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.
- Stubbs, M. and I. Barth (2003), 'Using recurrent phrases as text-type discriminators: a quantitative method and some findings', *Functions of language*, 10 (1): 61-104.
- Summers, D. (1996), 'Computer lexicography: the importance of representativeness in relation to frequency', in: J. Thomas and M. Short (eds.) *Using corpora for language research*. London: Longman. 260-266.
- Traugott, E. and B. Heine (eds.) (1991), *Approaches to grammaticalization*. 2 volumes. Amsterdam: John Benjamins.
- Wray, A. (2002), *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
- Zipf, G. K. (1945), 'The meaning-frequency relationship of words', *Journal of general psychology*, 33: 251-256.

Appendix

The 150 top 5-grams in the *BNC*, with the structure: preposition + determiner + noun + *of* + determiner (= *BNC* coding PRP ATO NN? PRF ATO), in descending frequency. The two allomorphs of the indefinite article are counted together.

at the end of the; by the end of the; as a result of the; in the middle of the; at the time of the; at the top of the; at the beginning of the; in the case of the; on the part of the; in the form of a(n); in the case of a(n); at the bottom of the; in the light of the; at the back of the; on the edge of the; for the rest of the; on the basis of the; at the end of a(n); in the context of the; in the centre of the; at the start of the; towards the end of the; at the foot of the; in the course of the; to the top of the; in the middle of a(n); in the hands of the; to the end of the; on both sides of the; in the direction of the; in the wake of the; at the heart of the; at the end of this; as a result of a(n); before the end of the; on either side of the; in the event of a(n); at the expense of the; until the end of the; with the rest of the; at the centre of the; on the basis of a(n); for the benefit of the; on the side of the; at the head of the; during the course of the; in the context of a(n); from the rest of the; at the turn of the; to the rest of the; in the absence of a(n); at the edge of the; on the back of the; in the course of a(n); in the history of the; for the purposes of the; under the terms of the; at the end of each; as a member of the; at the side of the; in the interests of the; as a result of this; in the face of the; on the floor of the; in the aftermath of the; in the back of the; to the back of the; at the front of the; in the heart of the; on the eve of the; under the auspices of the; in the form of the; to the edge of the; at the base of the; at the beginning of this; in the rest of the; with the exception of the; in the name of the; in the absence of any; to the bottom of

the; at the height of the; in the development of the; from the top of the; on the surface of the; in many parts of the; at the level of the; for the rest of his; in the eyes of the; in the shape of a(n); in the hands of a(n); under the control of the; at the end of his; to the right of the; to the left of the; on the top of the; to the attention of the; with the help of a(n); by the time of the; at the time of his; with the help of the; by the end of this; in the corner of the; to the side of the; after the end of the; in the course of his; from the end of the; as a consequence of the; at a meeting of the; in the event of the; on the banks of the; at the rear of the; in the life of the; on the day of the; since the end of the; to the needs of the; at the request of the; on the basis of their; at the centre of a(n); for the purposes of this; to the front of the; on the nature of the; for the duration of the; for the sake of the; for the use of the; on the face of the; to the centre of the; at the hands of the; on each side of the; on the edge of a(n); since the beginning of the; by the side of the; with the aid of a(n); to the north of the; about the nature of the; in this part of the; at the beginning of a(n); in the absence of the; from the back of the; on this side of the; in the face of a(n); as a result of his; into the hands of the; in the words of the; for the remainder of the; on the site of the; by the middle of the; in the presence of the; on the end of the; as a result of their; at the end of their.

This page intentionally left blank

The semantic properties of *going to*: distribution patterns in four subcorpora of the *British National Corpus*

Ylva Berglund* and Christopher Williams**

* University of Oxford

** University of Foggia

Abstract

In this paper the authors analyse how the intentional and predictive uses of the going to construction¹ can be seen to vary between different types of discourse, as found in BNC Baby, four one-million-word subcorpora from the British National Corpus. Selected collocational patterns of the construction are also examined.

As expected, results show that the overall frequency of the construction varies considerably between the text categories examined (newspapers, fiction, academic discourse and spoken conversation). Further interesting findings are made when this difference is put in relation to the predictive vs. intentional uses and the outcome of the collocational analyses. It is shown how the choice of main verb used relates to the distribution of intentional vs. predictive uses. Person and number are also taken into consideration as factors influencing the predictive vs. intentional ratios. Differences in semantic distribution patterns are also observed when gonna is used with respect to going to.

1. Introduction

As is well-known, and to the desperation of many non-native learners of English, the English language is particularly well-endowed with ways of expressing future time reference, including the present simple, the present progressive, *will* and *shall* with both simple and progressive forms, and a series of semi-modal constructions such as *to be to*, *to be about to* and *to be going to*. Traditionally, these constructions have been described and taught with a focus on their semantic properties; each verbal form of future time reference has its own meaning which in some way distinguishes it from every other, even if in some cases we seem to be dealing with nuances that almost defy description. What we are essentially concerned with here is to see whether corpus linguistics can help us to identify some of the semantic and pragmatic properties of one of these constructions: *going to*, a structure which, as Leech has observed (2003), is being increasingly used in spoken British English as well as in American English in general.²

Our approach to the subject was not so much corpus-driven in the sense used by Tognini Bonelli (2001) but was initially based on the works of several accredited grammarians, whose ideas about the semantics and pragmatics of the *going to* structure we wished to examine by analysing corpus data. While our results to date have not produced any major surprises, they do tend to suggest

some extremely interesting patterns of usage which, to the best of our knowledge, have not been highlighted by grammarians so far.

1.1 The *going to* construction

The salient semantic and pragmatic features of *going to* which are generally underlined by grammarians are

- its relatively informal style with respect to *will* (Huddleston and Pullum 2002: 211). The widespread use of *gonna* (as opposed to *going to*) in conversation is often a marker of informality; and it certainly is in written texts when spelt that way. We shall be looking into the *going to/gonna* question later in the paper;
- its dual meaning of ‘future fulfilment of present intention’ and ‘future result of present cause’ (Quirk et al. 1985), which have often been summed up as its intentional meaning and its predictive meaning;
- its tendency to be used to indicate the proximity of a future event unless there is a time adverbial or context indicating otherwise (Declerck 1991: 114). The fact that the structure is that of the present progressive form of the verb *to go* would seem to underline strongly its connection with the present (Williams 2002: 102).

Corpus-based, predominantly quantitative studies (for example Berglund 2005) have shown that the use of the construction varies considerably with a number of factors, genre being a major one.

In this study, our aim has been to see what correlation, if any, can be found between results obtained through a more traditional/qualitative analysis of the meaning of the *going to* construction based on the close reading of examples, and a corpus-based, primarily quantitative analysis. What can we learn about the use of the construction by looking at the semantic function as well as the distribution and collocation patterns? Can we find factors that show a relationship with the semantic criteria described by Quirk et al. (1985) and others by observing these patterns in different genres of English? Does the combined approach result in a clearer understanding of the linguistic issues at stake?

1.2 Corpus data and method of investigation

For our study, we opted to use *BNC Baby*, four subcorpora from the *British National Corpus* (BNC) (Burnard 2000). The subcorpora were compiled specifically to contain comparable amounts (one million words each) of texts from different types of genres. The corpora contain British English material with the bulk of the data from around 1990 (the earliest text in our subcorpora is from 1978, the latest from 1993). The *BNC* was automatically tagged with part-of-speech information (see Leech and Smith 2000). We used *WordSmith Tools*, Versions 3 and 4 (Scott) and *BNCweb* (see Hoffmann et al. 2002) to help extract and analyse data from the corpora.

By choosing these subcorpora, we can compare distribution across different types of text, including both written and spoken discourse. We are dealing with naturally-occurring present-day language and have the option of using the existing linguistic annotation. The tools we have opted to use allow us to quickly identify and count all instances of a particular item and, particularly in the case of *BNCweb*, offer advanced options for identifying collocational patterns.

We started our investigation by retrieving all instances of *going to* and *gonna* from the four subcorpora. All instances of the expressions which were used with (implicit or explicit) auxiliary verbs and infinitival marker *to* were extracted, while prepositional uses, such as *He is going to London*, were ignored for obvious reasons. We relied on a combination of manual inspection and automatic tagging to identify the relevant cases. Ambiguous or unclear instances were included. Both present and past tense forms (*is going to/was going to*) were extracted.

Once the instances of the *going to* construction were identified, these were analysed and classified according to the meaning of the structure in each instance (see Section 2 below). A number of parameters were taken into consideration in our attempt to distinguish between a predictive meaning and an intentional meaning. For example, a case such as *It's going to rain* is unequivocally predictive, whereas *I'm going to kill you* is clearly intentional. But what about *He's going to kill her*? Is this our prediction of his future behaviour or is it simply his intention? If we say *I bet he's going to kill her if he finds out*, then it would seem to be predictive, whereas *He says he's going to kill her because of what she's done* suggests an intentional reading.

As ever, context plays a crucial role in determining meaning. As we frequently discovered during our research, a lot of context is often required before it is possible to determine the meaning of the *going to* structure, and even then some cases were unclear or completely indeterminable.

In addition to the intentional and predictive classifications, we opted to use two additional categories: primarily but not exclusively intentional meaning (intention with some prediction) and predictive meaning with some intention. The semantic analysis was performed independently by both authors and the results were compared. In cases of disagreement, the instances were discussed and if agreement could not be reached, the native speaker analysis was selected.

2. Analysing the data

Figure 1 illustrates the quantitative distribution of the instances of *going to/gonna* across the four subcorpora in our study. As shown in the figure, the number of instances varies considerably between the genres, from less than one hundred instances per million words of running text in academic writing to almost 3,000 in spoken conversation. This gives a clear indication that the use of this construction varies with genre, something which has also been shown, for example, by Berglund (2005).

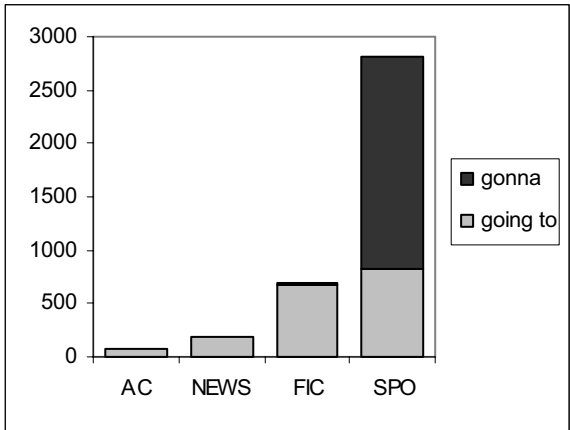


Figure 1: Frequency of *going to* and *gonna* in the four subcorpora.

The variant *gonna* was almost exclusively found in the spoken material, where it constitutes about 70% of all instances. When occasionally used in the written texts, it occurred in direct speech or quotes, as in:

The doorman revived, said, “Mah feets ain’t gonna stick roun’; to see mah body bein’ abused!” and scuttled off.
(Fiction. BNC GVL: 2878)

2.1 Predictive vs. intentional meaning in the four genres

Figure 2 illustrates the distribution of our analysis of the predictive (P) and intentional (I) uses of the construction, including the unclear or dual cases IP (intentional with some prediction) and PI (predictive with some intention). The results for the academic and news categories are based on the full sets of instances, while we had to opt to use a random sample from the much larger sets of instances in the fiction and spoken categories.

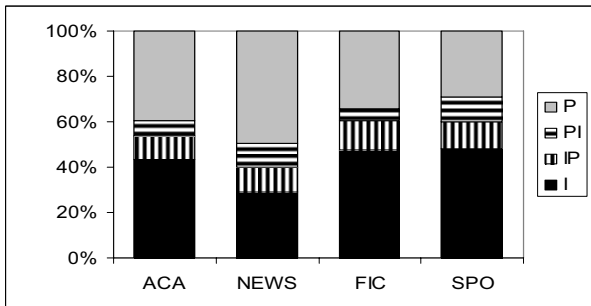


Figure 2: Distribution of predictive vs. intentional use across subcorpora.

If we examine the data relating to predictive vs. intentional meaning in relation to the four genres, we can observe that there is reasonably close agreement between the academic, fiction and spoken genres, all of which show some prevalence of intentional meaning, whereas the news genre alone reveals a marked preference for predictive meaning. About 20 per cent of all cases are given a dual reading (IP or PI). This is roughly the same across all four subcorpora.

2.2 Collocational Analyses

Firth (1957) suggested that a word shall be known by the company it keeps, that the meaning of an expression is influenced by the words surrounding it. In order to investigate how this correlates with our analysis of the meanings of *going to/gonna*, we examined some of the most frequent collocational patterns found with the structure.

2.2.1 Collocations with infinitival verbs

Like other semi-auxiliary constructions, *going to/gonna* is frequently used with infinitival verbs.³ When the total number of infinitives in the corpora is considered, it is found that the frequency is approximately the same for the academic and news genres (around 30,000) while it is higher in the spoken corpus (over 46,000) with the fiction genre being somewhere in the middle (37,000). We examined the collocations of the *going to* structure with the most frequently-used verbs, namely *be*, *have*, *do*, *go* and *happen*.⁴ The distribution is illustrated in Figure 3.

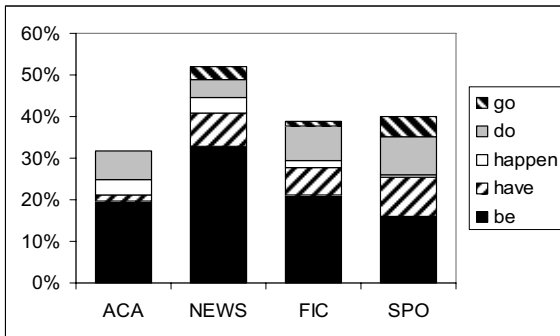


Figure 3: Proportions of most frequently-used verbs collocating with *going to/gonna*.

The most visible trends to emerge are:

- the clear predominance of *be* in all four genres;⁵
- the higher overall proportion of these verbs in the news genre;
- the very low proportion of *have* and *go* in the academic genre;
- the significantly high frequency of occurrences of *be* within the news genre in comparison with the other three genres.

Before examining these patterns in any detail, it must be remembered that the academic genre has a low incidence of the *going to* structure overall – only 76 instances in all. When collocational patterns with individual items are considered, the frequencies become very low indeed and should be interpreted with some caution. For that reason we will not go into the collocations in the academic genre in any detail. Patterns with *go* and *happen* must also be considered as mere indications since, although they are amongst the five most frequently-used verbs, they display low raw frequencies in collocation with the *gonna/going to* structure. Nevertheless, we have made one or two tentative hypotheses on the basis of the data available, though they will clearly require further verification by analysing much larger corpora.

2.2.2 Predictive vs. intentional meaning with the most frequently-used verbs

An obvious line of enquiry was to examine the predictive vs. intentional meaning of the instances of *going to/gonna* when occurring with these most frequently-used verbs. The results, as illustrated in Figure 4, revealed some very marked trends of usage, with the *going to/gonna* construction occurring predominantly with predictive meaning when used with *be*, *have* and *happen*, whereas instances collocating with *do* and *go* were found predominantly with intentional meaning.

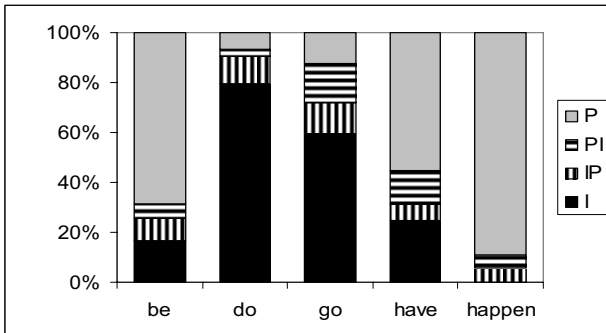


Figure 4: Predictive vs. intentional meaning with the most frequently-used verbs.

As illustrated above (Section 2.1), the *going to/gonna* structure occurs with a high proportion of predictive meaning in the news genre. The construction is also

frequently found collocating with infinitival *be* in that genre (Section 2.2.1). This means that the higher proportion of predictive meaning found in the news genre correlates with a higher proportion of the seemingly predominantly predictive verb collocation. As always, meaning is not the result of factors in isolation, but a tentative conclusion based on these results would be that *going to/gonna* used with the infinitival verb *be* will tend to render a predictive meaning.

2.2.3 Distribution patterns with conditionals

The use of the *going to* construction in conditional sentences has been studied by, amongst others, Declerck and Reed (2001). We decided to see what we could find out about the meaning and distribution of the construction in conditional sentences in our corpus.

Two clear patterns emerged: when used in the subordinate clause the *going to/gonna* construction displayed a marked tendency towards an intentional reading, e.g.

I used some big words on the way down. We're Italian now. Mumalissimo where are you bonjourdino. Oh Paul don't be silly! Go to bed *if you're going to be silly*. Go and learn your words for tomorrow (Spoken. BNC KD0: 3037-3041)

whereas when used in main clauses predictive usage clearly predominated, e.g.

But for the initiative of General Morillon and his transplanted European NATO staff, and of the individual units in UNPROFOR and various aid agencies, the whole operation would be hopelessly bogged down. Indeed, *it is going to be bogged down* unless the UN can get things moving with political and military action to make the peace plan work. (News. BNC K5C: 1281-1282)

However, it must be acknowledged that these results, although indicative, were based on a small number of instances.

2.2.4 Present/past forms

A frequent collocate of the *going to* structure is the auxiliary *be* – so frequent in fact that it is often considered as part of the construction. The instances we examined were usually found with an explicit auxiliary verb (*am/are/is* etc.), with the spoken genre displaying a somewhat lower proportion. This correlates with earlier findings (e.g. Berglund 2000a). As illustrated in Figure 5, the proportion of present tense realisations of the auxiliary is similar (70-75%) in all but the fiction genre (around 55%), which is hardly surprising given that most novels are conveyed in the past tense. This variation between the genres does not seem to correlate with the variation in predictive and intentional use of the *going to* construction.

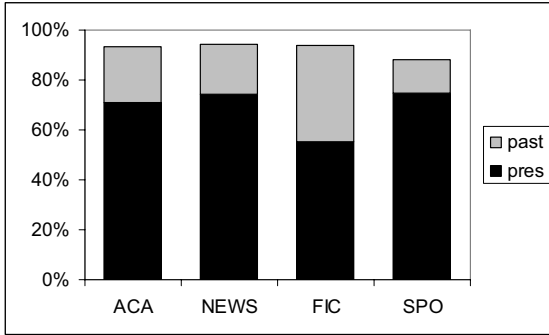


Figure 5: Present/past forms in the *going to* structure.

2.2.5 Distribution patterns according to person

As shown above, infinitival verbs are frequent collocates of the *going to* construction. Predictably, these verbs are almost exclusively found in the position immediately following the construction, while the auxiliary is found immediately preceding *going to/gonna* to a very large degree. The subject of the construction is usually found preceding *going to/gonna*, before or after the auxiliary. As shown previously (Berglund 2000b), *going to/gonna* is most often used with a personal pronoun subject. Figure 6 illustrates the person and number of subjects used with the examined instances of *going to/gonna* in this study.

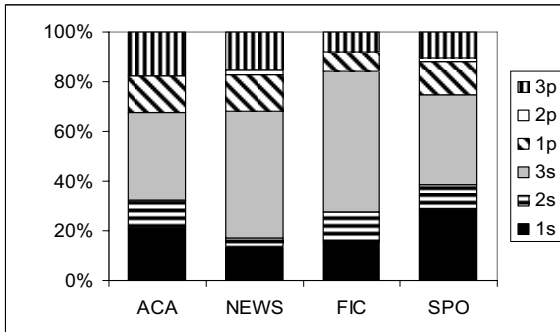


Figure 6: Subject of the *going to/gonna* structure.

As shown in the figure, the construction is most often found with a third person subject, with the singular being by far the more frequent. The proportion of first person subjects is lowest in the news genre and highest in the spoken corpus (it may be worth noting that the figures for second person subject should be interpreted with some caution where number is regarded. In some instances it is impossible to say whether *you* is directed at one or more persons).

Turning to the meaning which the instances of these subjects correlate with, we find that the distribution is very similar for singular and plural subjects in this case, as seen in Figure 7. There is a slightly higher proportion of unclear or double meaning cases for the plural subjects but this difference is not large enough to be considered here.

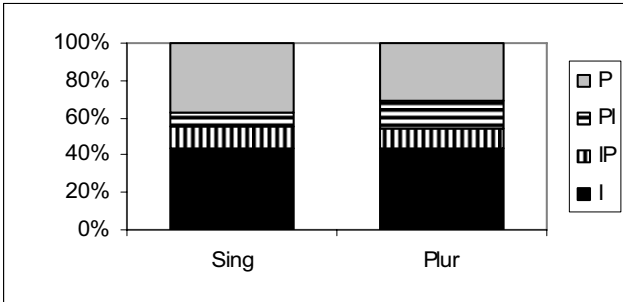


Figure 7: Meaning related to singular (S) and plural (P) subjects.

When looking at the person of the subject (first, second, third), there is a more marked difference between first and second person subjects on the one hand and third person subjects on the other, as shown in Figure 8.

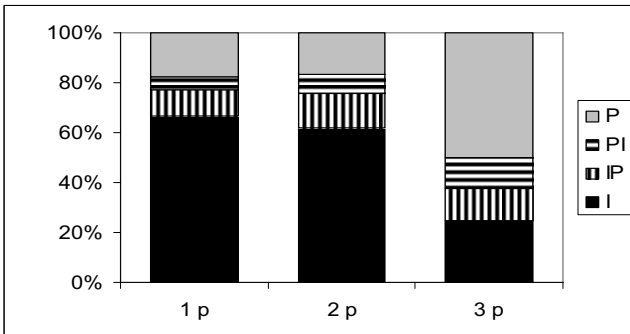


Figure 8: Meaning related to subject (first, second, third person).

Not surprisingly, the first and second person subjects are found to correlate primarily with an intentional reading while third person subjects are found mostly in a predictive use. This would suggest that texts with a high proportion of third person subjects could be expected to have a high proportion of predictive uses of the *going to* construction. Indeed, the genre in our study with the highest proportion of predictive use is the news corpus, which also has a larger proportion of third person subjects. (It is worth noticing that the number of

second person subjects is low, which discourages further analysis of that particular pattern at this time).

2.3 *Going to and gonna*

Previous studies (Berglund 2000a) have shown that *going to* and *gonna* are very similar in terms of collocation patterns: there is little difference between the variants where collocations with infinitival verbs, auxiliaries and subjects are concerned. However, if we examine distribution patterns on the basis of predictive vs. intentional meaning, some interesting results emerge. In our sample, *going to* is found in a more predominantly intentional use than *gonna*. Thus, while the proportion of unambiguously intentional cases outnumbered unambiguously predictive cases by about two to one with *going to*, the corresponding ratio with *gonna* is approximately four to three (see Figure 9). A detailed analysis of the reasons as to why this should be so, fascinating as they are, would take us beyond the scope of our paper. However, it is worth noting that the difference in distribution is quite a substantial one and certainly merits closer attention. But this can only be done exhaustively by taking into consideration the various alternative ways of conveying future meaning in English. For example, one might need to verify whether or not there is a greater use of the futurate present progressive form (as in *I'm catching the earlier train this evening*) in more informal registers of English which might partly account for the reduced frequency in intentional use of *gonna* and hence its corresponding increase in use in predictive contexts (see Römer 2005 for a detailed corpus-driven analysis of progressive forms, including futurate use, in contemporary British English). Moreover, the tendency to use the *will* future more in more formal contexts would probably help to account for the less frequent use of *going to* with predictive meaning and, vice versa, the higher frequency of predictive *gonna*.

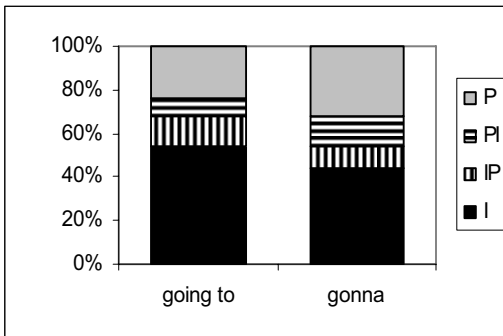


Figure 9: Meaning related to *going to* and *gonna*.

It is also worth noting the almost total absence in our data of the passive form with *gonna*, constituting less than one per cent of cases. With *going to* passive use is somewhat more frequent but still only accounts for 3.5% of the total. These lowly figures make it difficult to examine the difference between the variants further in this respect.

3. Concluding remarks: caveats and conclusions

We are fully aware that the data we have collected and analysed so far in our initial survey, as well as the methodological approach we have adopted, have a number of limitations which may diminish the degree of reliability of our findings. Two of the major drawbacks are related to well-known problems with the two approaches we have used (semantic analysis and quantitative corpus analysis respectively), and are as such hardly unique to our study:

- the relatively small number of instances that we have been able to analyse. As suggested by Sinclair (1991: 13), the results of a corpus-based study are only as good as the corpus on which it is based. If a corpus does not contain a sufficiently large number of a particular construction or pattern, conclusions based on that data can be but tentative. In our study, frequency has been an issue, in particular with regard to the academic subcorpus, where there were well under 100 instances of the *going to* construction in all. The solution to this problem would obviously be to analyse much larger corpora in order to arrive at more reliable figures relating to the various distribution patterns;
- the subjective nature in establishing whether the *going to* structure indicates predictive or intentional meaning, or possibly both simultaneously. Ways around this are difficult to find, although basing the analysis on the views of a large number of native speakers could be one way of making a study of this sort more objective.

Other problems worth mentioning are the uncertainty of the suitability or representativeness of the corpus as well as well-known issues related to the difficulty of interpreting the meaning of a text taken out of the context for which it was originally produced, be it spoken or written. Issues related to transcription of spoken data must be considered in any comparison between two variants such as *going to* and *gonna* (see Berglund 2005: 51-54).

In spite of these caveats, we still believe that our results display a number of interesting trends which are certainly worth exploring in more detail. To sum up the main points of our findings:

- there would appear to be a discernible differentiation in semantic distribution according to genre, the two major trends being the noticeably higher frequency of predictive usage in the newspaper subcorpus with

respect to the other three, and the tendency towards intentional usage in the spoken English and fiction subcorpora;

- collocation trends suggest a marked predominance of predictive meaning with *be*, whereas *do* and *go* tend more heavily towards an intentional reading;⁶
- in conditional sentences, subordinate clauses with *going to* tend towards an intentional reading, while main clauses with *going to* tend towards a predictive reading;
- indeterminate cases, i.e. when both a predictive and an intentional reading are possible, account for approximately 20% of all cases.

In addition to these linguistic results, we have found that using this dual approach, combining close semantic reading with quantitative methods, has been a fruitful way of approaching this topic. Of course, even corpus linguistics can only help us so far in delineating the semantic boundaries of the *going to* construction, and we shall probably have to leave to philosophers of language the ‘two for the price of one’ question. i.e. whether a speaker or writer may sometimes wish to convey both intentional and predictive meaning simultaneously, or whether this ‘fuzzy’ area of indeterminacy is merely the result of the addressee (or researcher) being incapable of deciphering the utterance with sufficient precision.

Notes

- 1 In this paper we shall generally refer to the construction as *going to* rather than *be going to*.
- 2 We wish to point out that there have been a number of other studies which have used corpora in relation to future time reference in English, though all from a different stance with respect to ours, for example Wekker (1976), Close (1977), Mindt (1991), Mair (1996), and Facchinetti (1998). See Bob Binnick’s extensive online bibliography for publications on the future tense in general (2002).
- 3 Römer (2005: 155-156) points out that five other verb forms, namely *expecting*, *hoping*, *meaning*, *trying* and *wanting* “also show this typical ‘TO BE V-ing to + infinitive pattern [...] It appears significant that the five verbs are semantically related in that they all refer to mental processes which refer to desires or intentions of differing degrees of intensity.”
- 4 It is interesting to compare the five most frequently used verbs with *going to* with the data provided by Römer (2005: 154) relating to the most frequently used verbs using the present progressive form with future time

reference, of which the first five are *coming*, *going*, *leaving*, *meeting*, and *sending*. Only the verb *go* is common to both sets of data.

- 5 *Be* can be found in 22% of all cases; this figure includes cases of the passive form as well as cases of the progressive form with *going to/gonna*, as in *He's going to be going up erm each weekend if this happens*, which constitute slightly under one per cent of the entire data (i.e. with all verbs).
- 6 Another area of research that requires further investigation is the distinction between statives and non-statives with *going to/gonna*. On the basis of a preliminary survey that we have carried out, non-statives outnumber statives by almost two to one, the former constituting 65% of all cases (*going to/gonna* combined). There is only a slight variation between *gonna* (62%) and *going to* (66%).

References

- Berglund, Y. (2000a), “‘You’re *gonna*, you’re not *going to*”: a corpus-based study of colligation and collocation patterns of the (*BE*) *going to* construction in Present-day spoken British English’, in: B. Lewandowska-Tomaszczyk and P. J. Melia (eds.) *PALC’99: Practical applications in language*. Frankfurt am Main: Peter Lang Verlag. 161-192.
- Berglund, Y. (2000b), ‘Utilising Present-day English corpora: a case study concerning expressions of future’, *ICAME journal*, 24: 25-63. Available from: <http://icame.uib.no/ij24/> [Accessed 13 October 2006].
- Berglund, Y. (2005), *Expressions of future in Present-day English: a corpus-based approach*. Uppsala: Acta Universitatis Upsaliensis. Available from: <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-5794> [Accessed 13 October 2006].
- Binnick, B. (2002), *Project on the bibliography of tense, verbal aspect, aktionsart, and related areas: future tense* [online]. Available from: <http://www.scar.utoronto.ca/~binnick/TENSE/futureTense.htm> [Accessed 13 October 2006].
- British National Corpus* [online]. Available from: <http://www.natcorp.ox.ac.uk/> [Accessed 13 October 2006].
- Burnard, L. (2000), *Reference Guide for the British National Corpus (World Edition)* [online]. Available from: <http://www.natcorp.ox.ac.uk/docs/userManual/> [Accessed 13 October 2006].
- Burnard, L. (2003), *Reference Guide for BNC-baby* [online]. Available from <http://www.natcorp.ox.ac.uk/corpus/baby/> [Accessed October 13, 2006].
- Close, R. A. (1977), ‘Some observations on the meaning and function of verb phrases having future time reference’, in: W. D. Bald and R. Ilson (eds.) *Studies in English usage: the resources of a present-day English corpus for linguistic analysis*. Frankfurt am Main: Peter Lang. 125-156.

- Declerck, R. (1991), *A comprehensive descriptive grammar of English*. Tokyo: Kaitakusha.
- Declerck, R. and S. Reed (2001), *Conditionals: a comprehensive empirical analysis*. Berlin/New York: Mouton de Gruyter.
- Facchinetti, R. (1998), 'Expressions of futurity in British Caribbean Creole', *ICAME journal*, 22: 7-22.
- Firth, J. R. (1957), 'A synopsis of linguistic theory 1930-1955', in: *Studies in linguistic analysis*. Oxford: Philological Society. 1-32.
- Hoffmann, S., Y. Berglund, D. Lee and N. Smith (2002), *The BNCweb manual* [online]. Available from: <http://homepage.mac.com/bncweb/manual/bncwebman-home.htm> [Accessed 13 October 2006].
- Huddleston, R. and G. K. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Leech, G. (2003), 'Modality on the move: the English modal auxiliaries 1961-1992', in: R. Facchinetti, M. Krug and F. R. Palmer (eds.) *Modality in contemporary English*. Berlin/New York: Mouton de Gruyter. 223-240.
- Leech, G. and N. Smith (2000), *Manual to accompany the British National Corpus (Version 2) with improved word-class tagging* [online]. Available from: http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm [Accessed 13 October 2006].
- Mair, C. (1996), 'The spread of the *going-to* future in written English', in: R. Hickey and S. Puppel (eds.) *Language history and linguistic modelling: a festschrift for Jacek Fisiak*, vol. II. Berlin: Mouton de Gruyter. 1537-1543.
- Mindt, D. (1991), 'Syntactic evidence for semantic distinctions in English', in: K. Aijmer and B. Altenberg (eds.) *English corpus linguistics*. London: Longman. 182-196.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.
- Römer, U. (2005), *Progressives, patterns, pedagogy: a corpus-driven approach to English progressive forms, functions, context and didactics*. Amsterdam/Philadelphia: John Benjamins.
- Scott, M. (nd), *Mike Scott's Web* [online]. Available from: <http://www.lexically.net/wordsmith/> [Accessed 13 October 2006].
- Sinclair, J. (1991), *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Tognini Bonelli, E. (2001), *Studies in corpus linguistics*. Amsterdam: John Benjamins.
- Wekker, H. C. (1976), *The expression of future time in contemporary British English*. Amsterdam: North-Holland.
- Williams, C. (2002), *Non-progressive and progressive aspect in English*. Fasano di Puglia: Schena Editore.

The superlative in spoken English

Claudia Claridge

University of Kiel

Abstract

The paper deals with the use of the superlative degree in spoken British English on the basis of the demographic part of the British National Corpus. The aspects investigated include the distribution of the morphological types (inflectional vs. periphrastic), the types of adjectives used in this construction and the syntax of the superlative (attributive, predicative and nominal use; determiner usage). Special attention is being paid to the semantics (relative, absolute, intensifying meanings) and the corresponding functions of the superlative, where it is noticeable that absolute and intensifying readings are much more common than expectable from the extant literature. Together with the usage of generalising modification structures, this points to the conclusion that the superlative may be less a means of factual comparison than rather a means for (often vague) evaluation and the expression of emotion.

1. Introduction

And worse I may be yet; the worst is not
So long as we can say 'This is the worst'.
(Shakespeare, *King Lear*, IV i. 27-28)

Superlatives can be used in various ways. The Shakespeare quotation, for example, indicates the potential of the superlative for saying more than is warranted, i.e. for overstatement. A possible penchant for expressing extremes or absolutes, whether contextually appropriate or not, is also reflected in Bolinger's (1977) remark that "the superlative can jump any adjective to the outer limits of its scale" (Bolinger 1977: 28). On the other hand, the use of superlatives in the wider context of comparison emphasises rather their relative meaning, i.e. a superlative predication can be located on any point of a given scale, by no means necessarily only the extreme ends. Thus, the superlative can have a clearly comparative and thus mainly factual use (e.g. a person can be *the tallest* of a specific group, measurable objectively in metres) or a looser and potentially evaluative/emotive interpretation (e.g. *the tiniest scratch*, without any comparative context and any attempt at objective measurement).

While the morphology of the superlative (and also the comparative) degree in actual language use has received some attention in the literature (e.g. Kytö and Romaine 1997), its semantic implications/interpretations have so far been mostly considered in theoretical (semantic/philosophical) studies (e.g. Veloudis 1998; Farkas and Kiss 2000; Sharvit and Stateva 2002). Rusiecki (1985) and Leech and

Culpeper (1997) are the only two studies based on authentic data commenting on the superlative *and* going beyond its morphology, but in neither of which is the superlative the main, or even at least a prominent, concern. The aim of the present study is thus to provide a more comprehensive account of the superlative in spoken English, paying attention to semantic and functional, but also to morphological and syntactic matters as well as to the potential interconnections of these various aspects.

2. Data and methodological considerations

The database chosen for this study is the *British National Corpus (BNC)*, specifically its spoken demographic part (to be referred to as *BNC-Sd*) – i.e. it is the everyday usage of the superlative that is being put in focus here. The 153 files of the *BNC-Sd* record the spontaneous conversations of 153 randomly sampled individuals from all over Britain, representing different social groupings with regard to gender, age and social class. The size of the individual files differs considerably and the corpus as a whole amounts to a total of 3,945,968 words. With its wide social scope and its emphasis on naturally occurring speech, the *BNC-Sd* seems suitable for yielding a realistic picture of the superlative as used in late 20th-century spoken British English.

In order to identify all occurring superlatives, all instances of *most* and *least* (tagged DT0 or AV0) and of the AJS tag¹ were sampled. Of the 982 hits for *most*, all those not modifying an adjective were excluded. *Most* accompanying an adverb (48 instances) carries superlative sense as well, but as there is no correlation of this use with *-est* adjectives (AJS) it will not be considered here. 34 AJS tags in fact also presented adverbial uses, most commonly the case with *best*, and were excluded as well. When *most* modifies nouns, it can be either the superlative of *many/much* or a proportional quantifier indicating somewhat vaguely “more than half, the majority” (cf. Huddleston and Pullum 2002: 1166). As these instances can be ambiguous between the two uses and hard to conclusively sort apart, they were disregarded in this study.² A similar procedure was followed for *least*, in which case the vast majority of instances (623 of 668) represented *at least*, leaving in the end only four *least* + adjective uses.

Searching for the AJS tag produced 1,841 instances, of which three were incorrectly tagged and excluded. Apart from the adverbial cases mentioned above, some other instances were removed as they represented written or quoted use. Furthermore, there are instances where the use or non-use of the superlative is not the result of a real option, whether in syntactic or lexical respect. Superlatives can occur as part of more or less frozen collocations or idioms such as *if the worst comes to the worst*, *make the best of*, *best before*, *nearest and dearest*, *at the latest* etc. (162 instances). These will also not be counted in the statistics given below. On the other hand, a cross-check with all words ending in *-est* revealed that adjectival superlatives were also found among other tag categories, namely general adverb (AV0), preposition (PRP) and proper noun

(NP0). Where these represented adjectival uses they were included in the data, which increased the whole by 79 instances.

3. Frequency and distribution

The procedures just described yielded the result of 1,973 true adjectival superlatives in the *BNC-Sd*, which comes to a frequency of 5.0 instances per 10,000 words. Compared to adjectival use in the *BNC-Sd* as such, the superlative is not very frequent, amounting to about 1.2% of all adjectives used (160,944 according to unambiguous tag usage). The comparative degree is more than double as frequent, with 5,615 instances (tags).³ On the whole, this points to the superlative not being a prominent means of marking (high) degree in conversation. In fact, Biber et al. (1999: 524) find that conversation has the least amount of inflected superlatives compared with the registers fiction, academic writing and news reportage, which last shows by far the greatest use; with periphrastic superlatives the situation is similar, except that here academic writing is in the lead. A concern for emphasising importance or even for a certain sensationalism in the case of news reportage and preference for certain types of (polysyllabic) adjectives in the case of academic writing, but not in casual conversation, may explain parts of Biber et al.'s result, but these do not provide a satisfactory reason for the relative rarity of the superlative in conversation. Perhaps a comparison with other degree/intensity-signifying items is helpful. *Very* occurs 6,523 times (16.5 per 10,000 words), i.e. is considerably more frequent than the superlative, whereas the much stronger *absolutely* and *extremely* are found only 671 times (1.7 per 10,000 words) and 56 times (0.14) respectively. That is, on the one hand, there are competing means to express (comparatively) higher degrees, cf. the following two examples, which might also have been expressed in a superlative form (*most favourite*, *most delicious*).

- (1) What's your **very favourite** food? (KBW 9998)
- (2) I just thought that was **absolutely delicious!** (KPU 381)

On the other hand, the figures seem to indicate that fairly extreme formulations may not be that common in spoken language.⁴ A search of the *FLOB Corpus*, representing written British English, confirmed this assumption for superlatives by revealing a much greater frequency of them in writing, namely 18 instances per 10,000 words. It may also be interesting to note here that Kytö and Romaine (1997: 332, 337) find a more even distribution of comparative and superlative degrees, and even a dominance of the latter, in their early modern English and late Middle English data as well as in the *Archer Corpus* up to 1850, all of which of course represent written language. This may further point to the possibility of a diachronic change, in fact decline, in the use of the superlative, which would merit further investigation.

As to the distribution, the attested superlatives are rather evenly spread through the corpus. They occur in 139 of the 153 files, and the fourteen files with no instances are all relatively short, from only 65 words to 5,165 words, so that the non-occurrence in those can be considered due to chance and insufficient data. Within the files, or rather conversations, the superlatives are again spread throughout the ongoing talk(s) and across various speakers. Smaller clusters and repetitions do occur, but are not the norm. But of course the normalised frequencies differ.

4. Characteristics of the superlative

4.1 Morphological types

As mentioned above, superlatives occur in a synthetic, inflectional form and in an analytic, periphrastic variant, the choice for which correlates largely with the type of adjective being compared. The inflectional variety has been found to be the more common one (Quirk et al. 1985; Rusiecki 1985; Biber et al. 1999), which holds also true for the present investigation, as can be seen in Table 1.

Table 1: Morphological superlative types.

	<i>periphrastic superlative</i>	<i>inflectional superlative</i>	<i>total</i>
<i>most</i> 257	261	1,712	1,973
<i>least</i> 4			
	13.2%	86.8%	100%

The difference inflectional/periphrastic is shifted considerably in favour of the former variant, even more clearly than in Kytö and Romaine's (1997: 335) data from a smaller subsection of the spoken *BNC*,⁵ where the relation is 73% to 27%.

145 different adjectives are found in the periphrastic superlative construction; the list is in the Appendix. Most of them appear only one to three times, but five types occur somewhat more often, namely *expensive* (18 instances), *important* (14), *beautiful* (10), *wonderful* (9), and *peculiar* (8). Nineteen of the adjectives could also, in some cases even more naturally, be used with the inflectional type, namely *fierce*, *nasty*, *odd*, *rude*, *rare*, *scary*, *ugly*, *weird*, *bitchy*, *strange*, *polite*, *common*, *pleasant*, *unfunny*, *friendly*, *likely*, *stupid*, *extreme*, and *little*. As regards the latter, the expected superlative forms would be *smallest* or *least*, but what is found here is *least little*, which seems to add emphasis, cf.:

- (3) Albert: Never were any good for sleeping on that!
 Rose: Now you know if my arm touched the back it would start it off.
 Albert: Yeah.
 Rose: **Least little thing**, even if

- Albert: You shouldn't be lifting it up and putting it at back of your head!
 Rose: Even the- I crumple it up like that
 Ada: Mm.
 Rose: and it irritates.
 Albert: You shouldn't have the arm up, you should have it down!
 (KB1 2530)

By using an additional word for the superlative, the force of the statement can be heightened (cf. Curme 1931: 500-501), which is what happens in (4).

- (4) No I, I, I sort of smelt Patrick's feet once and I thought if it was possible to faint on smells then I was going to, you know. I swear to God they're **the most nasty things** sort of like Aaron's (KP4 4372)

But not all of these alternative periphrastic forms add extra emphasis. Sometimes this form is chosen to produce an intensifying (not strictly superlative) use, which is only possible in an unambiguous way with the periphrastic type. *Odd, polite, strange* and *weird* provide examples for this point, e.g. (5). An additional point for the use of the periphrastic type in this example may also be the coordination with the form *courteous* with *most* doing service for both adjectives.

- (5) Oh yeah, yeah. Suit comes along and sits there sipping coffee. Takes about three quarters of an hour. I, I don't say that it was like being grilled by the KGB he was **most polite and courteous**. But he was going to find out where Paul was. He was going to find out. (KBF 6390)

A last point to be made on periphrastic superlatives concerns the form *most fun* in (6), seen as an adjectival form here (also tagged as such in the *BNC*).

- (6) that's what's **most fun** though, the talking innit Stew? (KCP 9288)

The form *funnest* appears to exist for some speakers, but probably not for all, and unwanted confusion with a wrongly produced form *funniest* might arise in the latter case.

Some forms occur with both periphrastic and inflectional forms in the *BNC-Sd*, namely *boring, nasty, strange, stupid, ugly, and little*. Thus, these also belong to the 109 different types attested with the inflectional superlative (cf. Appendix), a smaller range of types than in the case of the periphrastic group. However, the individual frequencies of occurrence are usually higher as a consequence. One type in particular is very frequent: 755 instances are found for the superlative of *good*, which amounts to 44% of all inflectional superlatives. The two next most frequent lag far behind with 109 (*big*) and 95 (*bad*) instances. Several adjectives here exhibit irregular forms of comparison, namely *good-best, bad-worst, far-furthest/farthest, old-eldest* (also regular *oldest*), and there is additionally the lexicalised form *foremost* (but still recognizable as superlative),

which has been included here. The irregular and lexicalised type *last* as a potential superlative of *late* has not been included in the data, as it has lost its superlative semantics. From the morphological point of view, the list of inflectional superlative adjective does not offer many surprises. The form *cackest* is parallel to *most fun* above in so far as an originally nominal form is used, although here *cackiest* would have been a viable option. The interesting participial forms *boringest* and *frustatedest* come from the same file of southern female teenagers talking among themselves, but not both from the same person. It might be that the unusual form is intended to produce some extra effect.

- (7) Kath: oh I love their platform shoes, I mean they're fucking rub they can't even walk in them.
 Anon: Mm.
 Roxy: No I just don't, I mean I don't understand what's ... like if you see who's going out with people it's always **the boringest people**, I've never seen any of them smile, laugh, talk, they just, they just seem like to be shells of people
 Kath: Mm. (KPH 660)

Double superlatives also occur in the present corpus, these are *most beautifullest*, *most easiest*, *most best* (once each), and five times *bestest* (also attested by Biber et al. 1999). The first occurs in a rather playful context, the second might reflect emphatic usage, while the last ones perhaps represent a counter-reaction to semantic bleaching.⁶ As mentioned above, *best* is very frequent, also in decidedly bleached idiomatic usages, and a considerable number of instances show uses of *best* with a reduced meaning, implying not much more than *good* (e.g. (8)) or acting as a marker in a narrative (9).

- (8) Carole: It's **best** not to put white in that colour.
 Pat: Not in green, no... they've got the white (KBH 2605)
- (9) A: But Margot, but Margot tells me that she's loaded. And she's a ring on her finger that's worth... what is it, Bernard, five grand or something?
 B: Oh aye.
 A: You know, so, I don't know. And then I, but **the best** of it was the other day ... we
 B: She's living near [...].
 A: the, the best, well **the best** of it was she was saying to us that er ... her uncle owns half of Gleneagles Hotel.
 B: Aye. (KE1 2625)

For some speakers, plain *best* might thus not carry enough superlative force any more and be in need of substitution.

4.2 Adjectives

A closer look at the set of adjectives occurring in superlative form reveals the fact that most adjectives in both groups are either inherently and primarily evaluative, have a second(ary), but established evaluative meaning component or can fairly easily be contextually filled with attitudinal meaning. Primarily evaluative examples are *abysmal, attractive, comfortable, difficult, exciting, gorgeous, horrid, outrageous, rude, sickening, wonderful, boring, crazy, lovely, posh, scummy, tasty, wise, funny, nasty, pretty, silly, dear, great, nice, good* and *bad*. Of the periphrastic group of adjectives 70% can be seen as primarily evaluative, a further 21% can be used either evaluatively or objectively, and only 9% denote a principally more objective quality (*absorbent, cost-effective, European, experienced, out of order, oversubscribed, populated, reactive, recent, religious, southern, unstable, used*) – but most of these latter can easily take on emotive connotations in particular contexts. About 50% of the inflecting adjectives are inherently evaluative, but most of the rest easily adopt an attitudinal component.

As we are here dealing with mono- or disyllabic adjectives of mostly native origin, most of them are polysemous and tend to have or acquire emotive overtones/connotations more readily than their polysyllabic Romance cousins. In many cases in the present data, the emotive potential of non-inherently evaluative adjectives is indeed made use of. Few adjectives in this group are exclusively factual, notably *eldest*, and *youngest*. In (10) *highest* is used with objective reference to amounts of money in a., whereas b. represents an already more evaluative use, going in the direction of *best*. Another potentially objective adjective of dimension, *longest*, is framed emotively in (11).

- (10) a. The **highest** my mum goes is forty pounds cos she doesn't think she should spend a lot of money on a pair of trainers that won't last.
(KPP 63)
- b. The church must make every effort to achieve the **highest** standards.
(KB0 1816)

- (11) That's the **longest** two minutes I've ever seen. (KBH 4240)

The superlative, or adjective comparison in general, involves the notion of degree, but not all adjectives are gradable as such. However, it has also been pointed out that gradability is not an either-or situation but can be contextually invoked (e.g. Huddleston and Pullum 2002: 531-532). The present group of adjectives is nevertheless characterised in the majority by inherent degree. A considerable number of those adjectives taking inflectional superlatives actually form part of antonymic pairs, or are located on the scale of gradable antonyms, e.g. *good-bad, big/large-small/little, cheap/dear-expensive, late-early, near-far, old-young, easy(-difficult), fast-slow, hard-soft, heavy-light, dark-light, high/deep-low, long-short, dirty-clean, cold-hot, poor-rich, happy-sad, weak-strong, ugly(-beautiful)*. This means that the notion of degree is very salient here.

The antonymic tendency is hardly present in the periphrastic group, but adjectives that are not clearly marked for degree are rare: possible candidates are *inopportune*, *European*, *out of order*, *two-faced* and also perhaps *incredible/unbelievable*. *European* in the geographic sense would usually be non-gradable, but if one uses the term in a more cultural sense it acquires gradability. *Two-faced* sounds like an either-or option, in particular as a concrete term (e.g. for the Janus head) but also as a character description; nevertheless, in (12) the superlative reflects the fact of grading the behaviour of two-faced people on a scale of offensiveness – which illustrates how context and communicative intention can render an adjective gradable.

- (12) I mean I tell you someone else who's **two faced**, Marcus is two faced. Marcus is one of **the most two faced people** I know, he's said some really nasty shit about me (KP6 750)

Example (13) is another case of non-prototypical gradability: *white* is strictly speaking not very amenable to degree, but here it is found within an idiom and it is the idiomatic meaning that is being graded.

- (13) But the background is that of fifty letters were sent out to a variety of people... er, lay and ordained... whatever that means! Erm... asking for permission to put their names forward... and from the, the few letters we had... the, this was [...] that we thought was **the whitest knights**... in terms of Geography... in terms of sex... in terms of... I wouldn't say intelligence, perhaps I should try to impress them! (KB0 2211)

It has been remarked on that inherently superlative or absolute adjectives, though prescriptively 'banned' from degree constructions, are not uncommonly graded in everyday usage (Biber et al. 1999: 526; Huddleston and Pullum 2002: 531-532),⁷ e.g. *most perfect*. The *BNC-Sd* does not yield any of the classic absolutes, such as *unique*, *complete*, or *impossible*; the only instance that belongs into that group is *extreme* as in (14).

- (14) Stuart: We [...] flying a glider or throw themselves of the bridge, if it doesn't work they die, somebody learns from that, you know, I know it, it really was a fucking [...]
 Mark: [...]
 Stuart: this, this has been **the most extreme** [...] of trial and error ever (KDA 6785)

This utterance takes place in the context of talking about inventions connected to the army/war and constitutes either an expression of admiration or of criticism (hard to decide with the given context), either of which is evaluative and may thus account for the strong form applied. Additionally, it is used in combination with the superlative-*of-N* construction, which may also serve emphatic purposes (cf. below). Even if strictly speaking absolute adjectives are almost absent, some

other items in the periphrastic group are indeed in themselves quite extreme, e.g. *abysmal*, *extraordinary*, *favourite*, *disastrous*, and *enormous*. With the exception of *extraordinary*, which is used in an intensifying construction (cf. below), the other cases are real superlatives with strengthened force. Within the field of size *enormous* is already very much in the upper reaches so that a further grading pushes thing to or even over the top – thereby transporting a strong attitudinal meaning.

- (15) Rachel: Well I'll never forget when we were younger, we had erm some Americans living across the road and they invited us er somehow [...] I was in the kitchen, they were going to give me lunch and she gave me **the most enormous**
- Michael: You probably barged your way in Rachel [...]
- Rachel: Probably. **Peanut butter and jelly sandwich** and I didn't know what to do with it because I certainly couldn't eat it. It was just appalling. I've never experienced anything like it (KPU 2719)

4.3 Syntax

There are three possible syntactic constructions in which superlatives can be used. The first construction is the attributive one, i.e. use as a pre-nominal modifier, as in *he married the youngest daughter* (KCJ 659) and the second is as predicative element, e.g. *your mind's most absorbent where you're up to the age of five* (KPU 2530). The third possibility, which is not always explicitly mentioned in treatments of the superlative, is the nominal use of the superlative phrase on its own, i.e. in the form that Huddleston and Pullum (2002: 415-416, 1054) term fused-head noun phrase. Examples are *Have I got the longest?* (KBW 5579) and *The yummiest was in the, is in the Sun* (KDY 920). With the periphrastic construction this usage is possible as well, e.g. *they'd sell you the most expensive because they'd got the most commission* (KP1 1328), but it is much rarer than with the inflectional type (cf. Table 2). There may be ambiguities between predicative use and fused-head nominal use; in order to avoid these as far as possible, I have only counted as nominal such as uses where the superlative is in subject or object position, or in an otherwise clearly nominal environment.

Table 2: Syntactic uses of the superlative.

	<i>periphrastic</i>	%	<i>inflectional</i>	%	<i>total</i>	%
attributive	184	70.5	1,147	67.1	1,331	67.5
predicative	71	27.2	300	17.5	371	18.8
nominal	4	1.5	218	12.7	222	11.2
unclear ⁸	2	0.8	47	2.7	49	2.5
<i>total</i>	261	100	1,712	100	1,973	100

The table shows that attributive use is by far the most common one, a result that confirms Rusiecki's (1985: 135-136) finding for the inflectional superlative, which was based on a smaller data set.⁹

Rusiecki (1985: 138) assumes that the preference for attributive position is due to the fact that this position makes for a simpler sentence structure. He makes this point by way of examples containing restrictive modification (e.g. *Among the boys in his class Jimmy is the tallest* vs. *Jimmy is the tallest boy in his class*), which, however, is the case for only about one third of the present data (cf. §6 below). Nevertheless, the integration of the adjective into a major nominal constituent leaves more flexibility for the filling of the other syntactic constituents, in particular the predication (which would otherwise be 'blocked' by the predicative adjective). Denser information packaging of this kind, however, is not a characteristic feature of casual conversation. The reason for the syntactic preference may additionally, or perhaps primarily, lie in adjective, and particularly superlative, functions and semantics. According to Croft (1991) and Taylor (2002: 178-179, 454-455) modification, and not predication, is the prototypical value of adjectives in the first place, thus making attributive use the primary one – and, one can add, as such also the one that can be expected to be more frequent.¹⁰ While the attributive adjective tends to encode stable, inherent properties (in Bolinger's [1967: 7] terms they fulfil the function of 'characterisation'), predicative adjectives are more likely to refer to a temporary property (Taylor 2002). That is, attributive adjectives define or restrict/delimit the reference of the noun more closely and thus form a more unitary relationship with it. The superlative in itself is a defining construction, as is also made obvious by the co-occurrence with definite determiners (*the*, possessive pronouns) contrasted with the ungrammaticality of indefinite determiners (*a*, *some*) in truly superlative constructions (e.g. Hawkins 1978, cf. §5 below). In semantic terms, if a certain property is specified in a high or at least comparatively higher degree by using the superlative, this property will be one that is (most) important or prominent at the time of speaking (i.e. in focus, topical) and one that defines the reference of the NP, thus is an inherent property in some sense, cf. the following two examples.

- (16) a. that's your **weakest area** (KST 187)
 b. we've had the **coldest year** (KCK 1006)

The meaning conveyed thus predisposes for attributive structure. Nominal uses of the superlative then elevate the property to *the* defining aspect, which can thus stand for the noun as such in the given context.

Determiner usage has already been mentioned, but this point merits a closer look. Definite determiners are the prototypical form occurring with the superlative, and they are indeed the most common ones, as Table 3 shows (definite article, demonstrative, possessive pronoun, genitive).

Table 3: Determiner use with superlative forms (in %).

	<i>def. article</i>	<i>indef. article</i>	<i>zero</i>	<i>demonstrative</i>	<i>poss. pronoun</i>	<i>genitive</i>	<i>quantifier</i>
inflectional	72.7	0.4	16.5	0.4	8.4	1.2	0.4
periphrastic	67.3	2.0	26.4	0.8	1.9	1.6	-

As regards the indefinite article, its use as such is surprising for the inflectional category, whereas for the periphrastic type its use is surprisingly low – given the fact that this form represents a well-established intensifying use (examples for both aspects, cf. below). The other notable fact about Table 3 are the relatively high frequencies for zero determiner. While zero is the expected form for intensifying predicative uses of the periphrastic form (and is indeed the most prominent form there), it also occurs in some cases with attributive periphrastic uses. With the inflectional type zero is fairly evenly split between attributive (121) and predicative (141) use. Some of the zero attributive types sound rather unusual, cf. (17) below, but they are too frequent to be regarded as errors. With predicative uses, one might assume a semantic difference between definite and zero determiner with the former being more clearly superlative and comparative in nature, but this assumption is not upheld by the data. (18), for example, occurs both with and without *the* in very short sequence.

(17) That’ll be easiest thing to do. (KB9 2743)

(18) which of Britain’s colonies is (the) most populated? (KDC 3 / KDC 5)

5. The semantics of the superlative

As the last section has already shaded into semantic matters, it is now time to look more closely at the meaning(s) of the superlative form. It will turn out that it has at least three distinct interpretations.

The superlative may be thought as indicating a very high degree, even the upper part of a scale and thus something rather absolute/extreme – which is also suggested by the sequence positive-comparative-superlative. However, it has been pointed out (by, e.g., Rusiecki 1985: 140 and Huddleston and Pullum 2002: 1161-1162) that superlatives do not necessarily work that way, but that their predication is relative to the compared set and does not make a factual statement about the presence or absence of a certain quality. *This one is the best of them all* does not necessarily mean that it is ‘good’, only that it is better/less bad than the other things under consideration. Example (19) makes the potential relativity of superlatives very clear by placing *best* within a frame of ‘badness’:

- (19) Anon: crap bands and I'm going but they're so untalented, sort it out and he's going...
 Cassie: What's he saying?
 Anon: he's just saying no but they're really good and I'm going but they're not, they're rubbish. I suppose [...]... Nirvana are **the best of a bad lot** by a long way.
 Cassie: Then why is it Nirvana never played... yeah I suppose you, I see what you mean yeah... (KP4 4107)

Thus, the superlative is relative both to some scale and to an intended set within which the comparison takes place. The relativity is most clear in certain uses where modification or the like specify the set (cf. below and §6), but it can also recede into the background in other contexts. It can also be argued that the sheer presence of the sequence positive-comparative-superlative creates a cognitive frame that can be exploited for less relative uses.

As just indicated, there are in fact various uses or interpretations of the superlative, which have been given labels such as relative, comparative, absolute, intensifying and quantificational – with the labels and their applications not fully agreed on. Farkas and Kiss (2000: 432), Sharvit and Stateva (2002: 454) and Huddleston and Pullum (2002: 1166-1167) define the absolute superlative in the following way: in sentences like *John climbed the highest mountain/Kim lives in the smallest house in England*, the relevant comparative sets are a contextually provided group, here all relevant mountains and all houses in England (i.e. $y\text{-est } X = X$ is the $y\text{-est}$ among all relevant X s). In the *mountain*-example, which is not restricted further like the second example (and assuming there is nothing else specified in the wider context), we find what Farkas and Kiss (437) term “‘absolute’ absolute superlative”, where the set can only correspond to the largest one possible and identifiable within our general world knowledge.

This last usage leads over to Fauconnier's (1975: 354-356) identification of “universal or modified existential quantified readings” of superlatives, which are similar to the effects of *any*, cf. *the faintest noise (= any noise) bothers my uncle*. In some sense, this usage or interpretation is even more absolute than the above readings, in so far as it produces a more extreme statement. According to Fauconnier, any superlative can be used this way provided that it represents the low point (cf. *faintest*) of the pragmatic scale applicable in a certain context (Fauconnier 1975: 356, 370; 1978: 293); also, there are certain licensing syntactic environments. The difference of the quantificational use to the ‘absolute absolute’ one is that the latter is definite and referential (indicating a specific mountain), while the former is not referential (Fauconnier 1980: 60, quoted in Veloudis 1998: 234), in so far, e.g., as there is no particular noise being referred to. Veloudis (1998: 233-234) therefore anchors this type of superlative in the epistemic world of the speaker. Thus, we have here a potential reading that is linked clearly to attitudinal and evaluative meanings (which links up with the type of adjective used, cf. 4.2. above).

In contrast to the absolute superlatives, the relative or comparative superlative expresses that *X* is the *y-est* among *Xs* as delimited by *Z*, where *Z* refers to, e.g. in the above-mentioned sentence *John climbed the highest mountain*, a group of mountaineers including John: in this case John climbed a higher mountain than anybody else in this group (Farkas and Kiss 2000: 432; cf. also Sharvit and Stateva 2002: 454, Huddleston and Pullum 2002: 1166-1167). Interestingly, Sharvit and Stateva (2002: 470) allow for the possibility that the absolute and comparative interpretations are not really two readings, but only different strategies for fixing the comparison set. Rusiecki's (1985: 137-139) understanding of the relative sense seems to be wider, including both the case just mentioned as well as the type *X is the y-est among all members of all X* explained above.

The final use to be found is the intensifying function (which Rusiecki (1985: 137, 140-142) calls "absolute"), where the superlative form equals the sequence *very, very/extremely* (or the like) + adjective. While Rusiecki (1985) and Huddleston and Pullum (2002: 1165) consider this to be possible freely only with periphrastic superlatives (with the exception of *dearest*), Quirk et al. (1985: 466) and Halliday (1994: 184) think that inflectional superlatives are also used in an intensifying way, e.g. the latter's example *he said the silliest things*. It is, however, only the periphrastic type that makes a structural difference between superlative proper (of whatever interpretation) and intensification, by employing the determiners *a(n)* or *some* (instead of *the*) in attributive and zero determiner in predicative position (Rusiecki 1985: 140); the inflectional type exhibits no such regular variation. But the regularity might not be that clear-cut with the periphrastic superlative either: Hawkins (1978: 236) points out that *the most X* can also be used in the intensifying sense, which means that *the most X* is ambiguous between two interpretations.

For the purposes of this study I will take a three-fold classification of readings as the basis, namely relative, absolute and intensifying. I will follow Rusiecki's wider interpretation of relative uses and, more precisely, regard as relative every instance which is accompanied by an explicit statement about the restricted set or standard of comparison, either within the sentence with the superlative form or in the surrounding co-text. This means that Huddleston and Pullum's absolute example *smallest house in England* (cf. above) is understood as relative here because of the presence of the restrictive prepositional phrase. The restriction makes clear that a limited frame is present within which a comparison takes place. Absolute, in contrast, means that no restriction is present in the co-text, producing an unqualified and thus potentially universal superlative. If the frame is the whole world/the speaker's or hearer's world knowledge (i.e. *S*'s epistemic world), the limits are either extremely high (the highest possibly imaginable) and/or vague, as well as often unverifiable. Unmodified *the highest mountain* may be taken to belong to a clear, though huge, set (all the mountains on earth) and to clearly refer to Mt Everest (given some world knowledge, or access to a printed encyclopedia),¹¹ but what about equally unmodified *the best movie* or *the most marvellous colours*? The set of movies, while finite at the time

of speaking (though of course always expanding), is too large for any single speaker to know completely and *best*, *marvellous* as evaluative epithets are not based on any objective, verifiable scales. It is thus hard to see any clear comparative frame and/or standard. Absolute superlatives are thus often referentially less well anchored, and they always carry connotations of extremity. Absolute uses shade into intensifying uses, the last group; there is no clear sharp boundary between these two interpretations. As in absolute uses there is no explicit restriction present in the co-text. Periphrastic cases where determiner usage (*a*, *some*, *zero*) clearly signals their status are unambiguous intensifying instances, but the superlatives with other formal characteristics, including inflectional superlatives, are accepted in intensifying use as well. In these instances, a replacement of the superlative form with the sequence intensifier + positive/plain adjective, e.g. *very*, *extremely*, *absolutely*, should yield a synonymous pair. It is noteworthy, however, that the superlative will be the stronger form on the expressiveness or intensity scale. The reason for the use of the superlative form instead of an intensifier must lie in its greater inherent intensity.

Let me discuss these three uses in more detail and illustrate them with examples, before I proceed to some statistics as to their distribution in the *BNC-Sd*. Example (20) is a clear relative instance, with *pub in the area* stating the frame of comparison and *customers* even supplying one comparative parameter.

- (20) we appear to be **the most consistent pub in the area**, with er customers and what have you. They all come in and tell us we're **the busiest** and I say well if we're **the busiest**, God help those that're **the quietest**. They cannot possibly survive. (KPA 1337)

Note that the frame supports both *the most consistent* in the same sentence, as well as *the busiest* and *the quietest* in the following sentence. A similar explicit instance, with the question form indicating a select group, is (21). *Most interesting* lacks a determiner, which above has been identified as a mark of intensifying use, but the situation is more complicated. Such cases are actually ambiguous between relative/absolute and intensifying meaning (cf. Huddleston and Pullum 2002: 1165), unless, like here, there is contextual evidence for one or the other or the synonymy test yields a clear result.

- (21) A.: Well okay, those of you who liked the Shakespeare better, for whatever reasons, even if it's just because you didn't get to grips with Johnson, *which of the plays*, first of all, erm... did you find **most interesting**?
 B.: Dunno.
 A.: Wi who's read, read Twelfth Night and found that one **most interesting**?
 Rebecca: I thought, I liked Twelfth Night cos it's funny. (KPV 7825)

Another clear relative example is (22), where the given comparative frame consists of only two items, a context where in fact the comparative form might even be a more natural choice.

- (22) its not only children's programme, the young chap said to the bird expert which is **the best** er *potatoes or bread?*, oh he says potatoes every time (KBB 5761)

A last example illustrates the use of an explicit standard of comparison. While the modification *in the world* may sound rather absolute, the restriction *lawns* as opposed to, e.g., membership numbers, represents a clearly factual comparison.

- (23) *In terms of lawns* it's the **biggest** club in the world. (KBB 6149)

All four instances so far have been examples of the type *the y-est among all relevant Xs* (above introduced under absolute), and this type is indeed the only relative use in the present data. This may be due to the fact that it is cognitively the simpler one: it requires comparison within one set/along one parameter only, while the second type *the y-est among Xs as delimited by Z* interlinks two sets in one comparison. Also, such types of comparison might not be required that often in casual conversation.

A special form of relative superlative are those instances where a number modifies the superlative and thus explicitly puts it within a larger set; this use also again highlights the general relativity of superlatives mentioned at the beginning of this chapter. Depending on how many medals of honour there are, the *sixth highest* in (24) is high or it may be not, but at any rate, all Danish medals constitute the comparative set.

- (24) Pu our own... Putney has been awarded the medal of honour, **sixth highest medal of honour**... for erm... services to the Danish industry and he's coming over and they're having a... a caviar and champagne reception at the Danish Embassy and now the Queen of Denmark... Queen Margarete is coming over. (KD0 4392)

Similar instances are *the eighth fastest time*, *second biggest cock*, *second oldest* (family member), *third best* (electronic product).

Another clear signal for relativity of the superlative is the presence of preceding phrases such *one/some of...*, *amongst the...*, as they make clear that a comparative set is involved. Modal adverbs such as *perhaps*, *probably* etc. also make a relative interpretation more likely, as they express doubt about the certainty of the classification and thus make an absolute interpretation impossible.

Some of the just quoted adjectives have meanings (in their non-abstract senses), in which they are easily measurable, e.g. *high*, *fast*, *big*, *old* (*oldest/eldest*); other instances are *cheap*, *deep*, *far*, *heavy*, *large*, *little*, *long*, *low*, *narrow*, *quick*, *slow*, *short*, *small*, *tall*, *warm* and *young*. These adjectives are

generally more likely to occur with relative interpretations, some even exclusively. In the case of *eldest* even the comparative frame is unchangeably set to the family.

A rather curious use is *a best friend* with indefinite article (cf. (25)), occurring twice in different conversations. The effect of indefinite *a* here seems to be a demotion from absolute status, from being ‘first’ among friends. That is, it anchors the description more firmly within a group of, probably, very good friends and thus combines relative with intensifying readings.

- (25) Josie: Wait a minute, I just gotta erm... what are you? **A best friend**.
 Kerry: Best, best, best friend! (KPG 2241)

It is interesting to note that it is possible to force rather extreme readings by further intensification of the superlative without giving up the relative nature of the superlative, as in the following two examples, where overall renovating/furnishing costs (26) and various house-buying deals (27) are the comparative sets.

- (26) That’s our *single* **biggest** cost, cos as I say, I had three quotes in total, and they were all sort of between thirteen and twenty three. (KBD 3492)
- (27) And he went through everything till he found out and it wa I didn’t have the abs the *absolute* **cheapest** one. (KB7 3278)

Absolute uses, on the other hand, leave everything much more open and have a wider scope. In (28) there is no restriction whatever; while the immediately preceding reference is to cashew nuts, using the ‘empty’ noun *things* leaves the potential reference wide open and *nicesst* can also refer to quite a few parameters. On the whole, this makes for a very generalising statement, which is largely produced by the absolute superlatives.

- (28) Why are the **dearest** things the **nicesst**? (KD0 1849)

The following example is tricky, as it could be a candidate for an intensifying interpretation, cf. *even extremely fine players*. However, the superlative allows certain semantic nuances which, I think, are intended here. The absolute interpretation here carries the sense of *the finest possible/imaginable player* and the fact that singular is used (much less likely in the intensifying paraphrase) adds the aspect of uniqueness. Absolute uses aim for the very top, which is not necessarily the case for intensifying uses.

(29) Don't ask me how he manages it but there is still an aura about Botham that intimidates even **the finest player**. Indian youngster... handled England's other bowlers with ease, he looked in tremendous form but when Botham came on and bowled him a harmless straight one, somehow managed to edge it into his pads and nearly played on. (KE2 5481)

(30) also aims for the top in some way, *best* follows after *absolutely fantastic* and *brilliant* and is meant to surpass those two. A substitution with intensifying *extremely good* would in this case rather be an anti-climax. It expresses a very emphatic extreme, cf. also the supporting use of *just*.

(30) Have you ever seen this [i.e. film]? It's absolutely fantastic, you've got to see it. It's brilliant it is just **the best**. (KNR 411)

A construction that is specialised for absolute use is the superlative-of-N construction as in (31), which implies *the most convincing of all possible manners*. Being an infrequent and marked construction, it has a greater impact that makes for emphasis (cf. Leech and Culpeper 1997: 369). Infrequent means that it occurs only 22 times in the *BNC-Sd*; but as it seems to have also a somewhat formal touch, it might be more frequent in context-governed speech and in writing.

(31) poor old William Roche who plays Ken Barlow in Coronation Street has been moaning and groaning and wincing and rising in **the most frightful and indeed the most convincing of manners**, [...] (KE2 3749)

In some cases the presence of modification can produce absolute readings rather than the more expected relative ones (cf. also §6 below). In (32) the post-modification *in the world* pushes the examples towards the extreme; notably, it is exactly the frame of reference that one has to substitute in 'absolute absolute' superlatives. However, as what is normally left unsaid is explicitly spelt out, it increases the force of the superlative. Incidentally, this is also an example of how superlatives can be used for the creation of irony.

(32) Chris: Quality control, I mean er... client satisfaction, things of that sort I'm not quite sure what he's into. But er being a local authority they're saying oh come on. Why make the effort.
 Lynda: I mean sadly that really was in a sense what making local authorities answerable was supposed to stamp out.
 Chris: Yes. And of course local government is **the most wonderful procrastinator in the world** isn't it (KBK 693)

As with relative uses, some adjectives seem to be more prone for the absolute function or even to almost specialise in it. These are the adjectives that lead to quantificational *any*-like reading (cf. above) and which do so at both ends of the

scale; Fauconnier (1975: 366) mentions *faint*, *remote*, *slight*, *small*, *tiny*, and *least*. Of those, *tiny* does not occur in this sense here, and while *small* does, it clearly prefers relative uses. Words behaving in a quantificational way are additionally *mere* and *foggy*. They specialise further in certain collocations, in particular with *idea*, or form idioms, e.g. *in the slightest*.

- (33) Bloody cigar packets. I said ooh we can't help you with them, I says I haven't **the faintest idea** of anybody that smokes them. (KCX 5738)
- (34) But he's just a bit irritating cos he... well... you know... I do, I don't understand how you go through a whole relationship and everything, and then come to end of it and not even feel **the slightest bit** (KBY 106)

The following three examples are typical cases of ambiguity between absolute and intensifying readings. No explicit frame of comparison and no structural indication is given. But in each case, an intensifying paraphrase yields an acceptable synonymous sentence: *most hideous plates/extremely hideous plates* (35), *an extremely dirty election* (36) and *a very stupid route* (37). Thus, quasi-intensifying use is at least likely, with the actual force of the examples lying somewhere between straightforward intensification and the absolute superlative.

- (35) Kath: I think it [i.e. brass, CC] must have been really fashionable at one point.
 Anon.: Yeah. We've got **the most hideous plates** in the drawers [...]
 Kath: Now you get them free from Esso and stuff
 Anon.: [...]
 Kath: if you send away fifty tokens. (KPH 74)
- (36) Alison: Do they're all gonna tell lies.
 Herbert: And they're gonna tell lies and it's gonna be a dirty lies.
 Alison: So it's not [laughing] [...] So you're not gonna pick up anything [...].
 Herbert: It's going to be **the dirtiest election**. They're going to mu throw mud at everybody. Everybody's going to be what's named and this that and the other. ... (KCF 261)
- (37) Anon.: This bus takes **the stupidest route**, it's [...]
 Meg: I know darling.
 Anon.: If it just went in a straight line it would take [...] fifteen minutes (KP4 3992)

If (35) had been intended as clearly and unambiguously intensifying, the indefinite article might of course have been used instead. This is the case in the following example, where the meaning conveyed is that of a 'perfectly acceptable present' and nothing more.

- (38) Brenda: Cos he said to me he said oh I suppose I'd better buy you a Christmas present and I said yeah buy me something that I'd use. He said what I said well if you're gonna me anything, buy me a tin of biscuits.
 Jean: Mm quite.
 Brenda: He said... tin of biscuits? I said
 Jean: I think that's **a most acceptable present**. ... It really is.
 (KBF 5711)
- (39) but I'm **most disgruntled** the way Geoffrey's getting on, he's not had a single parents evening this year, you know normally you go back to school, you have one September, October which is always too soon
 (KD6 3116)

A comparable predicative use, with zero determiner, is illustrated in (39). Adjectives in periphrastic comparison which occur exclusively in such clearly intensifying environments are *acceptable, ambiguous, amusing, attractive, delicious, disgruntled, envious, European, extraordinary, filling, friendly, generous, helpful, horrid, impressive, out of order, polite, resplendent, reverent, strange, surprised, ungrateful, unlike, unnatural, upset, and weird*, while those employed both in intensifying and superlative use are *beautiful, efficient, embarrassing, expensive, horrible, important, impressed, likely, obvious, odd, peculiar, rude, unusual, and wonderful*.

As stated above, intensification is impossible to clearly attribute in inflectional cases, but there are some interesting determiner usages. The example *a best friend* was interpreted above as approaching relative usage, but using the indefinite article can also be seen as an imitation of intensifying periphrastic instances. Both (40) and (41) seem very much to convey an intensifying intention.

- (40) Norrine: Erm [sniff] but Jan is one of the best organisers [...] that I know.
 Chris: Is she?
 Norrine: She covers [...] .She's terribly thorough.
 Chris: Mhm.
 Norrine: And her children are very, very well grounded.
 Chris: Are they? Yes.
 Norrine: You know, they, they really are. And that's **a greatest skill** in itself. (KBK 2013)
- (41) That was rather **a cheekiest thing** to do. (KBD 3388)

Such unusual determiner instances are very rare, but they show how the system can be extended and exploited to convey and make explicit certain semantic nuances.

A rather roundabout way of producing an intensification force might be the negation of a superlative, occurring in 69 instances. Statements such as *it was not the most sensible place to er have a group* (KC3 3906) or *Well he's not the most unfunny but he's not incredibly funny* (KP4 2180) are roughly equivalent to, respectively, *it was a rather/very stupid place* and *he's very/extremely boring*. Negation occurs with adjectives with mostly positive connotations, or at least neutral ones. Perhaps politeness plays a role here: it may be less face-threatening to use a positive term, scale it upwards and then deny it, thus still leaving a positive residue, than to use a more negative term in a straight-forward manner. Similarly, it is better in politeness terms to deny being the best driver (*I'm not the best driver in the world* (KDA 2454)) than explicitly praising oneself as being 'a very good driver', which is the intended meaning here.

Table 4 shows the distribution of absolute, relative and intensifying superlatives in the *BNC-Sd* data. Relative instances are fairly clear because of the criterion of contextual restriction, although it needs to be admitted that this group might be larger as extralinguistic context might have restricting force which is not visible in the corpus instances. Intensifying cases clearly present a problem. I have decided to count as intensifying anything, regardless of structure (cf. determiner use), that can be easily given a synonymous intensifying paraphrase – which procedure is of course subjective to a considerable degree. Thus, some of the figures in Table 4 should be seen not as hard-and-fast statistics, but as providing tendencies for the state of affairs in spontaneous language use.

Table 4: Interpretations of the superlative.

	<i>periphrastic</i>	%	<i>inflectional</i>	%	<i>total</i>
relative	96	37.1	846	50.8	942
absolute	67	25.8	738	44.3	805
intensifying	96	37.1	81	4.9	177
<i>total</i>	259	100	1665	100	

What is very notable in the table is that relative uses do not clearly dominate, which means that the basic function of superlatives, that of producing set comparisons, is not necessarily also the most common one. The majority of periphrastic superlative forms is absolute or intensifying, while these uses still come to almost half with inflectional types. These instances do not represent comparisons with an 'objective' basis, but are used in a semantically more loose, subjective and potentially extreme way. Most superlatives thus serve potentially evaluative purposes. If one sees this fact in connection with the mostly attitudinal adjectives which are used in this construction (cf. §4.2), the present data confirms the hypothesis of Leech and Culpeper's (1997: 368-369) that "the inflectional superlative can have a particular affective force" and is in its unqualified expression "emphatic" and shows that the same applies in even greater extent to the periphrastic variant.¹² A very important purpose of superlatives is thus not primarily comparison but evaluation and intensification.

The state of affairs just presented fits well to the fact that any substantiation or proof for the superlative statement is usually lacking. Objective comparisons might be backed up by evidence, but speakers' feelings and attitudes are simply postulated. Instances like the following two, with proof (of whatever kind) provided are thus rare indeed:

- (42) No, because I know ours are **the best**. *There's ninety eight percent meat in ours*. They are the best, they're made in Smithfield, Smithfield market, and they're the best ones yo I certainly haven't seen him come across any that are better. Even the Birds Eye ones, the pre-packed ones and what have you, frozen, are not as good as ours *and that's the customers that tell us that*. (KPA 1250)
- (43) Well *he's in the Guinness Book of Records* as **the best overall martial arts person** in the world (KPA 1303)

6. Modifying superlatives

The presence or absence of modification clearly restricting the comparison set was important for the definition of the interpretations of the superlative treated in §5, with modified examples being classified as relative. On the other hand, *the most wonderful procrastinator in the world* ((32) above) with prepositional post-modification was included under absolute usages. Therefore, it is necessary to look more closely at both the amount and the types of modification present in order to gauge its effects. Table 5 gives an overview of the types and frequencies, and it is obvious from a comparison with Table 4 that there is no perfect numerical overlap between relative usages and the distribution of modifications.

Table 5: Modifiers of superlatives.

	<i>inflectional</i>	<i>periphrastic</i>	<i>total</i>
relative clauses	194	28	222
determiners	136	8	144
prepositional phrases	167	53	220
other	37	4	41
<i>total</i>	534	93	627

On the whole, modification by prepositional phrases seems to be more prone to producing relative superlatives, as the clear majority of the phrases are fairly specific and restrictive. Thus one finds clearly defined references like *at the club*, *in [name of town]*, *in France*, *in his school class*, *in our/your year (school/university)*, *in the first division*, *in the pool*, *in this section*, *of all of the engines*, *of people employed in manufacturing*, *of the family*, *of them all*, *of the two/three of you*, *on the London market*, and *on the sheet*. All of these exactly delimit the relevant comparative set. On the other hand, there are some

prepositional phrases which specialise in the opposite effect; these have a rather wide and unspecific reference, thereby pushing the superlative towards the absolute sense or emphasising this further. Such instances are, e.g., *for miles around, in memory, in the world, of all* (44) and in particular *in the ... world* (45), (determiner variant: *the world's*), which, together with similar *on the face of the earth*, occurs 57 times. *Of all* in (44) does not in fact refer at all, but works rather as an intensifier.

- (44) Gavin: You glad that I cleaned up?
 Sue: Yes, I'm very glad! Because if I'd have come home and it'd been a mess I would have had a face **the longest of all!** ... I'd forgotten all about it actually Gav.
 Gavin: Had you? I even soaked the Christmas cake plate. (KC6 646)
- (45) Chris: What? ... What we getting?
 Lynne: A Fiat Uno.
 Chris: A Rolls Royce. Yeah, I thought, I thought we both agreed on Lotus. You want a Fiat Uno?
 Robert: Ori Orion.
 Chris: Are you serious? **The most unreliable**
 Lynne: Well
 Chris: **cars in the whole world!**
 Lynne: I'm only kidding! (KBM 178)
- (46) I mean it's not exactly, you know, *the world's safest Labour seat*, is it? (KBD 4957)

The fact that the modification in (46) produces a contradiction on the literal understanding – the world vs. (Labour seat in) Britain – shows how much superlative modification involving *world* has turned into an intensifier. To return to prepositional phrases proper, some of them can have both a precise and a loose use, correlating with relative or absolute use respectively. Examples are *round here, in my life, of your life, in this country, in England, in Britain, or in Europe*, which are potentially either somewhat vague (how far does *round here* extend?) or unverifiable (who has got complete overview over *in my life/of your life?*) or push the frame to some relevant limit of the speaker's and hearer's experience (e.g. *in this country*).

While (47) represents a relative use, (48) is an instance of absolute use: in (47) all the Swiss institutions within the political boundaries of Britain are at issue and being compared, whereas in (48) it is not all the males of the English population that are being compared – *in England* simply adds emphasis.

- (47) **The oldest Swiss institution** *in Britain*. (KCV 60)

- (48) Yeah. And er I said to Brian has John gone? He said... he's just revving up the car engine. And I [...] and I run to the door and... you'd just drove off. ... [laugh] You! It's no good coming over and looking at me like that Rick. You cheeky dog. You cheeky. [...] *the cheekiest dog in England*.
(KCL 2821)

Relative clause structures as modifiers are clearly specialised in widening or generalising the reference of the superlative, i.e. they are more likely to produce absolute or intensifying superlatives. These clauses have some features that make them useful for this function. There are, for example, clauses with rather indefinite time reference, i.e. either quasi-timeless present or present perfect indicating the complete/full experience of the speaker, e.g. *the best French cider I have drunk, the farthest north I've been, the nicest people I know, the longest conversation we have*. Many instances include modal verbs, pointing to a wide open potentiality, e.g. *the nicest start any one can possibly wish for, the best it can be, the finest green you can eat, the best you could expect*. The first of the foregoing examples also shows how items can be piled up to reach a greater, usually intensifying, effect, cf. *anyone, can, possibly* strengthening one another. Quite a number of clauses also contains the adverbial *ever* hinting at the absence of a limit to the superlative statement, e.g. *the funniest play ever written, the funniest movie I ever saw, the cleverest people you'll ever meet, the dirtiest election that there's ever been*. Instances (49) and (50) are combinations of *ever* with modal and present perfect respectively, again combining various features to increase the effect.

- (49) Tracey: I don't think Jane's got any erm... brains when it comes to men.
Linda: She's got no brains at all. She's got no brain! I mean... in school right
Tracey: Mm mm.
Linda: **the simplest test you could ever do...** she failed! She got three!
Tracey: [laugh] (KD2 656)
- (50) Margaret Oh, that's supposed to be a good shop!
Anon.: Ooh! Do you know that's... doughnut then is one of **these chewiest... most uninteresting doughnut**, I mean, *I've ever eaten!* (KST 353)

Ever also occurs as simple post-modifier (cf. (51)), in this use falling into the category 'other' in Table 5, which also includes post-modifying *around, here, there, this time/year, yet*, and pre-modifying *all time* – all of which attempt to go to the limits of the relevant context. *Ever* stands out, however, as being especially frequent with 104 individual or combined occurrences (= 16.6% of all modifications).

- (51) oh you missed Only Fools and Horses the other night, its **the funniest one ever** (KD6 830)

The remaining category in Table 5 is determiners, which includes all possessive pronouns, genitive forms of personal names and genitive paraphrases of post-modifying prepositional phrases (*Britain's, the world's*). While the latter work as illustrated above, possessives have the potential to clearly restrict the application of the superlative and thus predispose for relative use. *Your latest acquisition* (KPV 1665), for example, delimits the set of purchases to a certain buyer and *her youngest brother* (KBC 5342) refers to one specific family. However, the restricting force works in combination with the adjective and the noun used and is not always equally strong. As soon as these words are more evaluative, as in *your weakest area* (KST 187) or *my greatest fear* (KD0 258), the restricting power recedes while the meaning 'in the whole of a person's life-time/experience' (a huge outer limit) becomes more prominent, thus moving in a direction that makes the superlative more subjective. Subjectivity is another and the last aspect that should be highlighted here: many of the modifications discussed here, about 22% of all, involve explicit speaker self-reference (first pers. sg. pronouns). Superlatives are thus often overtly embedded in the ego-centric world-view of the speaker, which again speaks for an attitudinal, emotive use of superlative.

7. Conclusion

In contrast to the emphasis put in many treatments on the essentially relative nature of the superlative (e.g. Rusiecki 1985, Huddleston and Pullum 2002) even in "ordinary usage" (Jespersen 1949: VII, 392), the present study has shown that superlative forms have a high potential for uses other than factual comparison in everyday conversation. Intensifying and absolute interpretations are common, both of which represent the expression of high and of even extreme degrees. Modification structures used often have either universal scope or are anchored firmly in the subjective experience of the speaker. Lastly, the superlative adjectives themselves are largely of an evaluative nature. All this taken together means that the superlative needs to be interpreted as an important device in involved and expressive, emotive language styles. What remains somewhat puzzling in this respect is that it is so relatively infrequent in the present conversational data, a type of language that is usually classified as the involved type. A comparison of superlative usage in spoken versus written registers will thus certainly merit further attention.

Notes

- 1 Cf. Burnard (2000: 72-73) for the tagset used.

- 2 Huddleston and Pullum (2002: 1166) give the presence of or the possibility of inserting *the* as an indicator for the superlative sense. In the present data, only eight cases of 200 *most* + noun sequences are preceded by the determiner *the*, namely *the most: bottles, commission, kids, points, rain, sperm, votes* and *weight*.
- 3 The correctness of the tagging was not checked by me in these cases. Biber et al. (1999: 523) find the same approximate frequency relation between comparative and superlative degree in their corpus.
- 4 It needs to be kept in mind, however, that in the context of spoken language there are other means of conveying involvement and emotion, notably prosody, but also facial expression, gestures etc.
- 5 2,176,000 words of southern English speech from 75 *BNC-Sd* files.
- 6 One of them is produced by a seven-year old boy, however.
- 7 Cruse (1986: 216-217) mentions “implicit superlatives”, which should not be morphologically or lexically gradable. Judging from his example, the range adjectives he is thinking of seems to be wider than in the literature referred to above, cf. his examples *beautiful* and *enormous*.
- 8 The category ‘unclear’ includes those cases which are unclassifiable because of the usual features of spontaneous speech (e.g. false starts, interrupted speech). These examples could not be classified in any respect and will not appear at all in subsequent statistics.
- 9 His data consist of 13 files from the *Survey of English Usage*, amounting to 65,000 words.
- 10 A search on ten files of the *BNC-Sd* (KB0-KB9) revealed 4,560 instances of A+N sequences (attributive) vs. only 1,019 instances of *be*+A sequences (a typical predicative context). As the files in question contain altogether 13,900 adjectives (including ambiguous tags), the comparison can only be regarded as extremely tentative, however.
- 11 I nevertheless suppose that this does not exclude that unrestricted *the highest mountain* could be used also to refer to some other mountain, i.e. unrestricted absolute use leaves this option open.
- 12 In Leech and Culpeper’s data emphatic usage is much more common with inflectional than with periphrastic superlatives.

References

- Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman grammar of spoken and written English*. London: Longman.
- Bolinger, D. (1967), 'Adjectives in English: attribution and predication', *Lingua*, 18: 1-34.
- Bolinger, D. (1977), *Neutrality, norm and bias*. Bloomington: Indiana University Linguistics Club.
- Burnard, L. (ed.) (2000), *Reference guide for the British National Corpus* (world edition) (BNC CD-ROM). Oxford: Humanities Computing Unit of Oxford University.
- Croft, W. (1991), *Syntactic categories and grammatical relations*. Chicago: University of Chicago Press.
- Cruse, D. A. (1986), *Lexical semantics*. Cambridge: Cambridge University Press.
- Curme, G. (1931), *Syntax*. Boston: Heath & Co.
- Fauconnier, G. (1975), 'Pragmatic scales and logical structure', *Linguistic inquiry*, 6: 353-375.
- Fauconnier, G. (1978), 'Implication reversal in a natural language', in: F. Guenther and S. J. Schmidt (eds.) *Formal semantics and pragmatics for natural languages*. Dordrecht: Reidel. 289-301.
- Farkas, D. F. and K. E. Kiss (2000), 'On the comparative and absolute readings of superlatives', *Natural language and linguistic theory*, 18: 417-455.
- Halliday, M. A. K. (1994), *Introduction to functional grammar*. London: Arnold.
- Hawkins, J. A. (1978), *Definiteness and indefiniteness. A study in reference and grammaticality prediction*. London: Croom Helm; Atlantic Highlands: Humanities Press.
- Huddleston, R. and G. K. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Jespersen, O. (1949), *A modern English grammar on historical principles*. Part VII: *Syntax*. Copenhagen: Ejnar Munksgaard; London: George Allen and Unwin.
- Kytö, M. and S. Romaine (1997), 'Competing forms of adjective comparison in modern English: what could be *more quicker* and *easier* and *more effective*?', in: T. Nevalainen and L. Kahlas-Tarkka (eds.) *To explain the present. Studies in the changing English language in honour of Matti Rissanen*. Helsinki: Société Néophilologique. 329-352.
- Leech, G. and J. Culpeper (1997), 'The comparison of adjectives in recent British English', in: T. Nevalainen and L. Kahlas-Tarkka (eds.) *To explain the present. Studies in the changing English language in honour of Matti Rissanen*. Helsinki: Société Néophilologique. 353-373.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.
- Rusiecki, J. (1985), *Adjectives and comparison in English. A semantic study*. London/New York: Longman.

- Sharvit, Y. and P. Stateva (2002), 'Superlative expressions, context and focus', *Linguistics and philosophy*, 25: 453-504.
- Taylor, J. (2002), *Cognitive grammar*. Oxford: Oxford University Press.
- Veloudis, I. (1998), "'Quantifying" superlatives and *Homo Sapiens*', *Journal of semantics*, 15: 215-237.

Appendix

Adjectives with periphrastic comparison:

absorbent, abundant, abysmal, acceptable, accurate, aggravating, alarming, amazing, ambiguous, amusing, annoying, appealing, attractive, awful, awkward, beautiful, best, bitchy, boring, brilliant, civilised, cockiest, comfortable, common, complex, confident, considerate, consistent, convenient, convincing, cost-effective, dangerous, delicious, delightful, difficult, diletante, disagreeing, disastrous, disgraceful, disgruntled, easiest, effective, efficient, elegant, embarrassing, enormous, envious, European, exciting, expensive, experienced, extraordinary, extreme, fabulous, famous, fantastic, fascinating, favourite, fierce, filling, friendly, frightful, frustrating, fun, generous, gorgeous, hateful, helpful, hideous, honest, horrible, horrid, horrific, important, impressed, impressive, incredible, inopportune, interesting, likely, little, lucrative, marvellous, nasty, necessary, obvious, odd, out of order, outrageous, oversubscribed, passionate, peaceful, peculiar, pleasant, polite, popular, populated, powerful, rare, reactive, recent, refreshing, relaxing, reliable, religious, resplendent, restrictive, reverent, revolting, romantic, rude, scary, sensible, serious, sickening, snobbish, southern, stimulating, strange, stupid, successful, surprised, surprising, terrible, two-faced, ugly, unbelievable, uncomfortable, unfunny, ungrateful, uninteresting, unlike, unnatural, unreliable, unstable, unusual, upset, used, useful, valuable, vulnerable, weird, wonderful

Adjectives with inflectional comparison:

1 instance: *bitter, boring, cack, chewy, cold, crazy, cushy, daft, deep, dry, fit, foremost, frustrated, grassy, green, healthy, horny, lovely, lucky, mere, mighty, naff, narrow, noisy, poor, posh, pushy, raunchy, remote, rich, sad, scummy, shitty, soft, straight, strange, stupid, tall, tasty, tough, unlucky, weird, white, wide, wise, yummy*

2-10 instances: *brainy, brave, bright, busy, cheeky, clean, clever, close, dark, dirty, early, far, foggy, funny, happy, hot, kind, light, little, mean, nasty, new, pervy, pretty, pure, quick, quiet, rough, safe, short, silly, simple, slow, strong, thick, tight, tiny, ugly, warm, weak*

148 *Claudia Claridge*

11-50 instances: *dear, easy, faint, fast, fine, great, hard, heavy, high, large, late, long, low, nice, slight, small, young*

51-100 instances: *bad, cheap, near, old*

101+ instances: *big*

700+ instances: *good*

Occurring in both groups: *boring, little, nasty, strange, stupid, ugly*

Semantically-based queries with a joint *BNC/WordNet* database

Mark Davies

Brigham Young University, Provo, Utah

Abstract

The British National Corpus (BNC) contains a wealth of data about the frequency and distribution of words and phrases in many different registers of English, yet, via the standard interface, there is not explicit way of investigating the semantic relationship between words. On the other hand, WordNet contains detailed hierarchies about the semantic relation between hundreds of thousands of lexical items, but it has very limited information about the frequency and distribution of these words. My project employs relational databases to join together these two resources, and allow advanced semantically-based queries of the BNC. These include queries that show the relative frequency of all of the synonyms for a given word, which hyponyms (more specific words) or meronyms (part of a whole) of a particular word are more common in the BNC, and all of the specific phrases that express more general semantic concepts.

1. Introduction

Certainly two of the most important linguistic resources for the study of English are the *British National Corpus* (www.natcorp.ox.ac.uk) and *WordNet* (www.cogsci.princeton.edu/~wn/) (Aston and Burnard 1998; Burnard 2002 for the *BNC*; Fellbaum 1998 for *WordNet*). As is well known, the *BNC* contains 100 million words of text from a wide variety of registers, while *WordNet* contains a semantically-organised hierarchical database of hundreds of thousands of lexical relations. Considered from another point of view, we find that although the *BNC* contains detailed frequency and distributional information on lexical items in English, it provides very limited information on the semantic relationships between items (primarily because this was not part of its original scope). *WordNet*, on the other hand, provides precisely the opposite information. It contains a wealth of information on semantic relationships and hierarchies, but says relatively little about frequency or distributional facts regarding these items.

What would be ideal, of course, is to join these two resources together. One can easily imagine the benefits of using the frequency and collocational information from the *BNC*, and joining this together with the extensive semantic hierarchies encoded in *WordNet*. For example, a user could find any of the following:

- which of thousands of different nouns occur with several different synonyms of *bad*, such as *bad idea*, *foul mood*, *wicked witch*, or *evil eye*;

- which terms for parts of the body or parts of a car are most common, including the frequency in different registers;
- which hyponyms (more specific words) of [to walk] are the most common, again with the possibility of comparing frequency counts across registers.

As one can readily appreciate, such queries would be useful for native speakers of English, but they would be even more useful for learners of English, who do not have native-speaker intuitions regarding the relative frequency of given lexical items and collocations. For example, a non-native speaker may not know that *foul mood* and *wicked witch* are quite a bit more common than *severe mood* and *evil witch*. Likewise, a beginning learner would likely have encountered *to walk*, but would have little idea of more specific words for [to walk] like *stagger*, *stroll*, *clomp*, and *pussyfoot*, and even less idea of their frequency and distribution across different registers.

Ideally, each of these types of searches could be done via a simple web-based interface and involve just one simple, quick query. This paper outlines how such as project has in fact been carried out, using relational databases that link together the *BNC* and *WordNet*. To actually use the online corpus, the reader is referred to <http://view.byu.edu/>, which is available free of charge.

2. The *BNC* in relational database form

As mentioned, in order to be able to link *WordNet* and the *BNC* together, we first have to get the *BNC* into a relational database. We first start with the 4500+ raw text files that compose the *BNC*, which have a linear structure like the following:

- (1) [...] <w PRP>within <w AV0>even <w AT0>a <w AJ0>small <w NN1>group <w PRF>of <w NN0>people <w PRP>at <w NN1>work<c PUN>, <w EX0>there <w VM0>will <w VBI>be [...]

We first strip out all of the headers, and we then place each word/POS pair on separate lines. The final files contain more than 100 million rows of data like the following (in this case we have placed one set of rows to the side of the other to save space):

Table 1: Vertical structure for the *BNC*.

<i>ROW</i>	<i>POS</i>	<i>WORD</i>		<i>ROW</i>	<i>POS</i>	<i>WORD</i>
50891887	<w PRP>	within		50891893	<w PRP>	at
50891888	<w AV0>	even		50891894	<w NN1>	work
50891889	<w AT0>	a		50891895	<c PUN>	,
50891890	<w AJ0>	small		50891896	<w EX0>	there
50891891	<w NN1>	group		50891897	<w VM0>	will
50891892	<w PRF>	of		50891898	<w VBI>	be
50891893	<w NN0>	people				

Next we import the 100+ million rows of text into MS SQL Server, creating a table of 100+ million rows. Each row contains just three columns: a sequential [ID] number to identify each successive row, a [word] column, and a [POS] column (as in the table above). We then run an SQL command which – for each row in the database – finds the next six words and places these in additional columns of the table. The database then contains 100+ million successive seven word sequences, as in the following table:

Table 2: *N*-grams table.

ID	WORD1	POS1	WORD2	POS2	WORD3	POS3	...	WORD6	POS6	WORD7	POS7
50891887	within	PRP	even	AV0	a	AT0	...	of	PRF	people	NN0
50891888	even	AV0	a	AT0	small	AJ0	...	people	NN0	at	PRP
50891889	a	AT0	small	AJ0	group	NN1	...	at	PRP	work	NN1
50891890	small	AJ0	group	NN1	of	PRF	...	work	NN1	,	PUN
50891891	group	NN1	of	PRF	people	NN0	...	,	PUN	there	EX0
50891892	of	PRF	people	NN0	at	PRP	...	there	EX0	will	VM0
50891893	people	NN0	at	PRP	work	NN1	...	will	VM0	be	VBI

This main [7-grams] table can then be converted to specific [x-gram] tables, by collapsing identical rows and placing the number of identical rows as a new column in the database. For example, the following table shows a small fragment of the [3-grams] table with some of the entries for the lemma [break] as a verb. The table contains the [word], [lemma], and [POS] for each unique three word string, and the first column indicates how many times that exact string occurs in the *BNC*.

Table 3: Example of 3-grams where lem1 = BREAK and word2 = THE.

FREQ	WORD1	POS1	LEM1	WORD2	POS2	LEM2	WORD3	POS3	LEM3
106	breaking	VVG	break	the	AT0	the	law	NN1	law
98	break	VVI	break	the	AT0	the	law	NN1	law
56	broke	VVD	break	the	AT0	the	silence	NN1	silence
53	break	VVI	break	the	AT0	the	news	NN1	news
46	broke	VVD	break	the	AT0	the	news	NN1	news
40	break	VVI	break	the	AT0	the	deadlock	NN1	Deadlock
24	broken	VVN	break	the	AT0	the	law	NN1	Law
23	break	VVI	break	the	AT0	the	habit	NN1	Habit

As one might imagine, these tables – although much smaller than the main 100+ million row [7-grams] table – are still quite large. For example, there are more than 800,000 rows in the [1-grams] table (i.e. 800,000+ unique types in the *BNC*). This increases to 4.9 million rows of unique [2-grams], 9.4 million rows of unique [3-grams] and 8.2 million rows for unique [4-grams]. In order to have smaller tables (and therefore faster data retrieval), we limit all [2-gram] and

higher tables to just those *n*-grams that occur two times or more in the *BNC*. If we include even the *n*-grams that occur just once, the tables are much larger – about 11 million rows for [2-grams] and 40 million rows for [3-grams].

At this point it may be profitable to briefly compare our approach to previous work with *n*-grams databases of large corpora. Perhaps the first corpus to use this architecture was the 100 million word *Corpus del Español* (www.corpusdelespanol.org), which I created from 2001/2002. Based on this architecture, there was the subsequent ‘Phrases in English’ (PIE) website and database that has been created by Bill Fletcher (<http://pie.usna.edu>). The PIE site is based on the *BNC*, and it allows users to search for sequences of words and/or POS tags, and then see the original context for the matching strings.

Users of our website (<http://view.byu.edu/>) use simple query syntax to search for the frequency and distribution of strings. For example, to search for all examples of *break* + the + noun, they simply input the following into the search form:

(2) break the [nn*]

(A drop-down list also inserts the part of speech tag, for those who are not familiar with the *BNC* tagset). The web-based script then queries the [3-grams] table and returns the following hits. As with the PIE site, users can click on each of these to see the phrase in context.

Table 4: BREAK (v) THE [noun].

<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>
law	280	back	16	strike	10	chocolate	5
news	134	world	15	hold	9	connection	5
silence	98	power	14	skin	9	contact	5
rule	82	glass	13	club	8	course	5
deadlock	58	seal	13	impasse	8	health	5
ice	50	camel	12	monotony	8	bond	4
spell	41	code	12	cycle	7	bread	4
habit	35	contract	12	peace	7	company	4
mould	35	door	11	heart	6	consensus	4
chain	34	pattern	11	line	6	diet	4
surface	32	story	11	stillness	6	egg	4
link	26	window	11	stranglehold	6	engagement	4
bank	25	agreement	10	tension	6	fall	4
record	20	journey	10	term	6	chocolate	5

There is one important difference between the PIE site and the database that we have created, however, and this difference deals with coverage. As we have mentioned, our databases contain *all n*-grams. The PIE database, on the other hand, is limited to just those *n*-grams that occur three times or more. This may not appear to be overly significant, but in terms of *n*-gram frequency it is quite important. By increasing the coverage to all *n*-grams, we create databases that are

roughly four to five times as large as those that contain only the strings that occur three times or more. In other words, the PIE loses approximately 75-80% of all *n*-grams by including only those strings that occur just three times or more.

A clear example of the importance of including even less common *n*-grams is the following table. This table shows a portion of the [3-grams] for the phrase [BREAK] [THE] [NOUN]. Note the interesting strings like ‘break the cohesion’, ‘break the mood’, and ‘break the Sabbath’. Each of these potentially adds some insight into the meaning of ‘to break’, and yet with a more limited database like that of the PIE site, we would not be aware of such strings.

Table 5: Less frequent strings for [BREAK (v) THE [noun]].

<i>activity, ban, barrier, bone, boredom, bound, boundary, circle, cohesion, concentration, conspiracy, country, court, day, dependency, director, enchantment, enigma, filter, fish, force, government, jar, kiss, marriage, mood, organisation, post, pound, promise, regulation, resistance, routine, sabbath, sequence, sound, stone, sunday, thread, top, train, trust, un, union, wheel, wicket, will, yalta</i>
--

3. WordNet in relational database form

Creating the *WordNet* database is somewhat easier than the *BNC* database. There is already a version in relational database form at <http://wordnet2sql.infocity.cjb.net/>. While this is in MySQL and DB/2 format, it was easily ported over to MS SQL Server, the database used in our project. The database contains the following tables, and it is the interaction of these tables that provides the power behind the SQL queries.

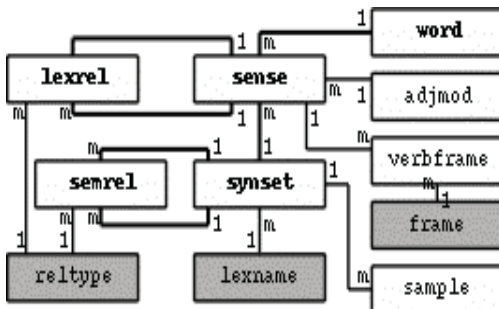


Figure 1: *WordNet* tables.

One of the central database tables is the [sense] table, which contains more than 200,000 entries containing many different ‘synsets’ or word senses of tens of thousands of words. For example, the following table is just a partial listing of the many different entries for [beat], for several different parts of speech:

Table 6: Synsets of [BEAT] (partial list).

<i>ID</i>	<i>WORD</i>	<i>POS</i>	<i>SYNSET</i>	<i>LEXFILE</i>
103065	beat	A	(informal) very tired	adj.all
2475	beat	N	a stroke or blow	noun.act
1401	beat	N	the act of beating to windward	noun.act
26292	beat	N	a regular rate of repetition	noun.attribute
35537	beat	N	(prosody) the accent in a metrical foot of verse	noun. communication
76162	beat	V	wear out completely	verb.body
78744	beat	V	be a mystery or bewildering to	verb.cognition
80925	beat	V	beat through cleverness and wit	verb.competition
80912	beat	V	come out better in a competition, race, or conflict	verb.competition
82481	beat	V	give a beating to	verb.contact
82490	beat	V	hit repeatedly	verb.contact

In the following sections, we will see how this basic ‘synsets’ table can be used (at times in conjunction with other tables) to find and display to the user all of the synonyms, hypernyms, hyponyms, meronyms, and holonyms for each of the different word senses of a given word, and then how this semantic information is merged with other tables to show the frequency of the semantic concepts in the *BNC*.

4. Basic synonym queries via the web-based interface

Perhaps the most basic use of the *WordNet* databases is to find all of the synonyms for the different senses of a given lexical item. Let us continue with the example of *beat* given above. In order to access the *WordNet* information and look for synonyms, the user would enter the following into the search form:

(3) [=beat].[v*]

The first part of the query string ([=beat]) indicates that we are searching for synonyms of *beat*, while the second part ([v*]) indicates that we are interested just in *beat* as a verb. The following is a screenshot of the search interface:

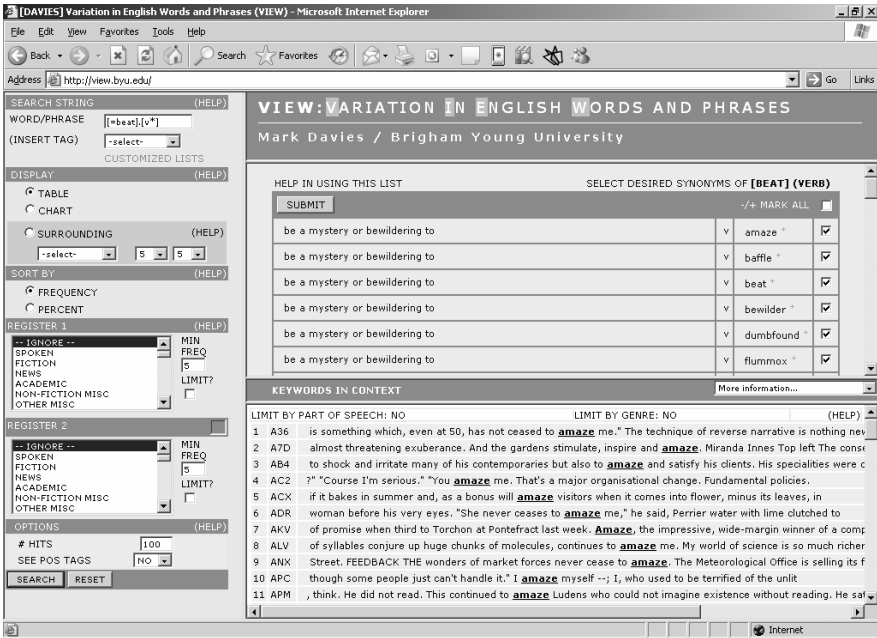


Figure 2: Screenshot of the VIEW/BNC interface.

The user then sees a listing of all of the different synsets for *beat* as a verb, along with all of the other verbs that share each of these meanings. The following is a more detailed view of some of the entries from this results set, which represents the upper right frame seen above:

Table 7: Partial results list for [beat] as a verb.

SUBMIT	+/- MARK ALL		<input type="checkbox"/>
be a mystery or bewildering to	v	amaze	<input type="checkbox"/>
be a mystery or bewildering to	v	baffle	<input type="checkbox"/>
be a mystery or bewildering to	v	Beat	<input type="checkbox"/>
be a mystery or bewildering to	v	bewilder	<input type="checkbox"/>
.....
be superior	v	beat	<input type="checkbox"/>
beat through cleverness and wit	v	beat	<input type="checkbox"/>
beat through cleverness and wit	v	circumvent	<input type="checkbox"/>
come out better in a competition, race, or conflict	v	trounce	<input type="checkbox"/>
come out better in a competition, race, or conflict	v	vanquish	<input type="checkbox"/>

If users want to focus in on a particular meaning, they can select just that one entry on the web-based form. For example, suppose that the user is primarily interested in the meaning of *beat* expressing the concept [to be a mystery or bewildering to], as is "*it beats me why she says such things*". After selecting just this one entry, the web-based script then finds all of the other words that express

this concept (“*it beats/puzzles/bewilders/perplexes me*”, etc), as well as the frequency of each of these words in the *BNC*. After clicking on any of these words in the lower frame, the user then sees a KWIC display for that word (with the correct part of speech) in the *BNC* (note that in this table there are reduced left and right contexts to fit this printed table):

Table 8: KWIC display for words displayed in Figure 2 above.

<i>TEXT</i>	<i>LEFT</i>	<i>WORD</i>	<i>RIGHT</i>
K5L	children respect ? Respect ? they	puzzle,	what's that ? The head teacher
EW7	— an article which would greatly	puzzle	dog fanciers who had turned to the
CE9	thin air ; it was only then did they	puzzle	and wonder if the dusk had conned
H7H	There were no Pommés Anna to	puzzle	him, but would he find the croûtons
CAB	and stared up at the ceiling trying to	puzzle	it out. Finally he gave up and
FYV	plenty of opportunity to do — to	puzzle	at it, I mean. I puzzle a lot,
BML	the narrative is intended to	puzzle	(is he doing it or dreaming it ?),
CKY	Service lists just 50,000. Dead dogs	puzzle	archaeologists The largest dog

To this point we have considered how the query takes place, from the point of view of the end user. Now let us go somewhat deeper and consider briefly how the query is processed in terms of the underlying *WordNet* and *BNC* databases. The following is one of the key SQL commands, which generates the table seen in the middle frame of Figure 2 – all of the synonyms for each of the synsets of the desired word (in our case [beat] as a [verb]):

- (4) `select distinct s1, IDs1, w1, c1 from [sense] where IDs1 in (select IDs1 from[sense] where w1 in ('beat') and c1 = 'v') order by ID pos1 asc, lexfile2 asc`

Because each synset (‘meaning’) has a unique ID, the SQL command find all of the other words in the synset database that also have the ID belonging to one of the synsets of [beat]. For example, the following table lists the other lexical items that have ID #78744, which belongs to the synset that expressed the concept [be a mystery or bewildering to]:

Table 9: Lexical items in a synset.

<i>IDI</i>	<i>WORD</i>	<i>ID</i>	<i>SYNSET</i>
112659	amaze	78744	be a mystery or bewildering to
23583	baffle	78744	be a mystery or bewildering to
2187	beat	78744	be a mystery or bewildering to
112656	bewilder	78744	be a mystery or bewildering to
112660	dumbfound	78744	be a mystery or bewildering to
112657	flummox	78744	be a mystery or bewildering to
112253	mystify	78744	be a mystery or bewildering to
112658	nonplus	78744	be a mystery or bewildering to
111840	perplex	78744	be a mystery or bewildering to

5094	pose	78744	be a mystery or bewildering to
32936	puzzle	78744	be a mystery or bewildering to
111301	stupefy	78744	be a mystery or bewildering to
112655	vex	78744	be a mystery or bewildering to

These thirteen words are the ones that appear in the first synset of Table 7. If the user selects this synset, then a subsequent web-based script stores these thirteen words in a temporary database. The script then retrieves these words and inserts them into a query that searches for the frequency of each of these words in the main *BNC* 1-grams table. This provides the output for the table in the lower frame of Figure 2. A final script then finds KWIC-formatted output from the *BNC* for any word selected by the user, as in Table 8 above.

5. Synonym-based collocations

The preceding example demonstrates one of the more basic uses of the *BNC/WordNet* database. In this case, we use *WordNet* to find all of the synonyms of a given word, and then use this output to find the frequency of each of these words in the *BNC*. However, this could have also been done manually. In other words, we could have done the following: go to the main *WordNet* site at Princeton (<http://www.cogsci.princeton.edu/cgi-bin/webwn>), enter in [beat], select the synset [be a mystery or bewildering to], see which other words belong to this synset, copy and paste the first word from the list into a *BNC* interface (e.g. <http://sara.natcorp.ox.ac.uk/lookup.html>), look at the KWIC display, go to the next of the thirteen words in the list, go through the same process, and so on through each of the thirteen words. With our approach, however, we can carry out this process for any number of synonyms of a given word in just one or two simple steps.

With more complex queries, the advantage of our approach becomes even more pronounced. For example, suppose that a user wants to see all of the collocations involving a synonym of [wicked] followed by a noun. The user simply enters the following into the search form:

(5) [=wicked] [nn*]

This searches for all synonyms of [wicked] from *WordNet*, followed by any noun. In less than four seconds, the user then sees something similar to the following. (The format on the web interface is somewhat different from the abbreviated listing shown here).

Table 10: Synonyms of [wicked] + NOUN.

<i>PHRASE</i>	<i>FREQ</i>	<i>PHRASE</i>	<i>FREQ</i>
disgusting thing	16	severe shortage	26
disgusting way	5	severe weather	43
distasteful species	5	severe winter	49
evil empire	10	terrible accident	26
evil eye	24	terrible blow	13
evil influence	9	terrible danger	14
foul language	29	terrible feeling	19
foul mood	21	terrible mistake	48
foul play	72	terrible shock	41
foul temper	14	wicked grin	6
severe blow	44	wicked people	13
severe burn	28	wicked thing	27
severe damage	59	wicked way	14
severe drought	30	wicked witch	12
severe illness	23		

Such collocational data can be very useful for a language learner, who is probably unsure of the precise semantic range of each adjective. The type of listing given above, which shows the most common nouns with each of the adjectives, can easily permit the language learner to make inferences about the semantic differences between each of the competing adjectives. For example, s/he would see that *severe illness* occurs but *wicked illness* does not, and that *terrible mistake* is common, whereas *foul mistake* is not.

In terms of text processing, we should note that this type of query would be quite cumbersome with the standard *BNC* interface, and would even be quite difficult with another interface like *BNCweb* or the ‘Phrases in English’ sites described above, both of which allow searches by part of speech. Again, the difficulty is due to the fact that successive queries would have to be carried out for each synonym of [bad], and the output from each query would then have to be collated together.

6. Related concepts through strings of synonyms

Perhaps an even better example of the power of the database is one that contains a string of synonyms, which express a given semantic concept. For example, suppose that one wants to look for all synonyms of [large] followed by all synonyms of [amount], such as *large sum*, *big amount*, *great measure* and *large total*. Users would simply enter the following into the search form:

(6) [=large] [=amount]

Within two seconds, the user then sees the frequency for each of these phrases from the *BNC*, such as the following. (This is again an abbreviated listing of what would be seen via the web interface).

Table 11: RESULTS: [large] + [amount].

	<i>PHRASE</i>	<i>FREQUENCY</i>
1	large (a) amount (n)	793
2	large (a) quantity (n)	488
3	large (a) sum (n)	373
4	large (a) measure (n)	146
5	great (a) quantity (n)	88
6	great (a) amount (n)	71
7	great (a) measure (n)	45
8	great (a) sum (n)	10
9	big (a) amount (n)	6
10	big (a) sum (n)	6
11	large (a) total (n)	2

Imagine if this semantically-based query were carried out with a standard interface. The user would have to input separately each of the synonyms of [large] with each of the synonyms of [amount], which might be on the order of 60 different combinations [10 x 6]. With our interface, it is just one simple query.

7. Semantic hierarchies – hypernyms and hyponyms

WordNet allows us to study much more than just synonyms. For example, we can find all of the words related to a given word whose meaning is more specific [hyponym] or more general [hypernym]. With our joint *BNC/WordNet* database, even if the list contains 40-50 words we can quickly find the frequency for all of these words in the *BNC* with just one simple query.

The query syntax is quite simple. The following two symbols are used to extract hyponym and hypernym entries from *WordNet*, and enter them in as part of the *BNC* query:

- (7) [$<$ word_x] more specific words relating to [word_x] (hyponyms)
 [$>$ word_x] more general words relating to [word_x] (hypernyms)

For example, suppose that a language learner wants to find out more specific ways of expressing the concept [to walk], involving words like *amble*, *prowl*, *saunter*, and *skulk*, as well as the frequency of each of these words. In order to find these more specific verbs, the user would enter the following into the search interface:

- (8) [$<$ walk].[v*]

(Where [*<walk*] indicates that we are searching for words that have a more narrow [*<*] meaning than [*walk*], which are verbs [*v**]). Behind the scenes, the web-based script would invoke this following SQL query:

- (9) `select distinct top 100 w1,c1,s1 from x_sense where IDs1 in (select s1.IDs1 from x_sense as s1 left join x_semrel as r on s1.IDs1 = r.IDs1 left join x_sense as s2 on r.IDs2 = s2.IDs1 where r.relation = 'hypernym' and s2.w1 = 'walk' and s2.c1 = 'v') order by s1 asc`

The data would then be output to the user, and s/he would see a listing like the following:

Table 12: More specific terms for [*walk*] (partial listing).

<i>SYNSET</i>	<i>WORDS</i>
take a walk	promenade, stroll
take a walk for one's health or to aid digestion, as after a meal	constitutionalise
to go stealthily or furtively	creep, mouse, pussyfoot, sneak, steal
to walk with a lofty proud gait, often in an attempt to impress others	cock, prance, ruffle, sashay, strut, swagger
tread or stomp heavily or roughly	trample, tread
walk (informal)	foot, hoof, hoof it, leg it
walk about	Ambulate
walk as if unable to control one's movements	careen, keel, lurch, reel, stagger, swag
walk by dragging one's feet	scuffle, shamble, shuffle
walk heavily and firmly, as when weary, or through mud	footslog, pad, plod, slog, tramp, trudge
walk impeded by some physical limitation or injury	Hitch, hobble, limp
walk in one's sleep	sleepwalk, somnambulate
walk leisurely	amble, mosey
walk leisurely and with no apparent aim	saunter, stroll
walk on and flatten	tramp down, trample, tread down
walk on one's toes	tip, tippytoe, tiptoe
walk or tramp about	shlep, traipse
walk ostentatiously	exhibit, march, parade
walk stealthily	Slink
walk stiffly	Stalk

As before, the user simply selects the synsets that are of interest, and then clicks on 'Submit' to see the frequency of each of these words in the *BNC*. The following listing, for example, shows the relative frequency of more specific verbs relating to [*walk*]:

Table 13: BNC frequency counts for hyponyms of [walk] (partial listing).

WORD	FREQ	WORD	FREQ	WORD	FREQ	WORD	FREQ
march	1765	trample	286	keel	108	scuffle	46
creep	1481	trudge	257	slink	104	leg it (p)	42
stride	1050	plod	212	stomp	104	bumble	40
stumble	1011	tramp	190	waddle	97	sleepwalk	26
tread	893	saunter	185	slouch	96	hoof	21
stroll	770	foot	183	hike	89	promenade	19
shuffle	687	prowl	182	prance	82	sashay	19
stagger	672	hobble	172	shamble	77	perambulate	13
sneak	452	amble	167	skulk	69	careen	10
trot	450	flounder	160	flounce	60	mosey	10
stalk	422	stump	156	swagger	59	dodder	9
reel	406	totter	155	slog	56	swag	8
lurch	397	tiptoe	151	traipse	51	leg (n) it (p)	6
limp	374	strut	146	toe	49	pussyfoot	4
parade	372	lumber	135	toddle	47	lollop	2

The following is another example of finding more specific lexical items to express a given concept. Suppose that a language learner has forgotten the word [aquarium], but does know that it refers to some type of [tank]. S/he would simply enter the following into the search form:

(10) [<tank]

With one more click, s/he would see the frequency for each of these items:

Table 14: More specific terms for [tank], with frequency counts (partial listing).

SYNSET	WORDS	FREQ
(German) an armored vehicle or tank	panzer	25
a heater and storage tank to supply heated water	hot-water heater	–
	hot-water tank	–
	water heater	79
a large gas-tight spherical or cylindrical tank for holding gas to be used as fuel	gas holder	5
	gasometer	12
a tank for holding gasoline to supply a vehicle	gas tank	10
	gasoline tank	–
a tank or pool or bowl filled with water for keeping live fish and underwater animals	aquarium	988
	fish tank	95
	marine museum	–
	vivarium	7

8. Finding the frequency of more and less specific words

In addition to looking for more specific or general words relating to a specific concept, it is also possible to use *WordNet* to find the parts of a given whole (meronyms) or see what larger unit a given word is part of (holonyms). This may be useful for language learners as well. Many textbooks will simply give a list of words in a given semantic domain ('body', 'parts of a house'), with a simple one word equivalence from the first language, e.g. [waist/*cinturón*]. The language learner, however, may want to move beyond such a simple list to see a more detailed definition, as well as see which words mean essentially the same thing, and see the frequency of competing items. Again, this would be quite easy with our joint *BNC/WordNet* database.

The query syntax is quite straightforward. The following two symbols are used to extract hyponym and hypernym entries from WordNet, and enter them in as part of the *BNC* query:

- (11) [*@whole* *x*] words that are a parts of [*whole* *x*] (meronyms)
 [*&part* *x*] words that contain [*part* *x*] (holonyms)

For example, to find the meronyms (parts of whole) for the whole = [body], the user would enter the following into the search form:

- (12) [*@body*]

The user then sees a display similar to the following:

Table 15: [BODY] ("Includes part") (partial listing).

<i>SYNSEMANTIC</i>	<i>WORDS</i>
(anatomy) a muscular partition separating the abdominal and thoracic cavities	diaphragm, midriff
a ball-and-socket joint between the head of the humerus and a cavity of the scapula	articulation humeri, shoulder, shoulder joint
a human limb	arm, leg
a protruding abdomen	belly, paunch
any of several muscles of the trunk	serratus, serratus muscles
the angle formed by the inner sides of the legs where they join the human trunk	Crotch, fork
the system of glands that produce endocrine secretions that help to control bodily metabolic activity	endocrine system
the fleshy part of the human body that you sit on	arse, ass, backside, behind, bottom, bum, buns, butt, buttocks, can, derriere, fanny, fundament, hind end, hindquarters, keister, nates, posterior, prat, rear, rear end, rump, seat, stern, tail, tail end, tooshie

As before, the users select the desired synsets, and can then see they frequency for each item in the *BNC*, as in the following:

Table 16: *BNC* frequency counts for parts of body (partial listing).

<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>
head	37940	waist	1381	torso	251	paunch	56
body	31421	trunk	1123	heads	214	posterior	44
back	20130	rear	1107	fanny	206	butts	42
arm	18933	fork	1018	diaphragm	157	waistline	42
leg	11176	belly	903	prat	119	pressure point	37
shoulder	8203	cavity	534	bum	510	caput	27
middle	5758	bum	510	butt	421	fundament	11
neck	5615	butt	421	haunch	106	tush	11
chest	3743	stern	371	cervix	98	derriere	7
cheek	3228	ass	350	backs	92	serratus	5
tail	3139	backside	311	crotch	73	midsection	4
stomach	2983	rump	311	behind	70	dorsum	2
hip	1791	abdomen	296	midriff	67	necks	2
can	1429	buttock	270	thorax	66	shoulder joint	2

If the user wishes to focus on just one of the synsets, s/he can easily do so. For example, the user may select just the synset [the fleshy part of the human body that you sit on], and would then see just the following:

Table 17: Terms for [the fleshy part of the human body that you sit on].

<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>	<i>WORD</i>	<i>FREQ</i>
seat	10464	stern	371	prat	119
bottom	5848	ass	350	behind	70
tail	3139	backside	311	posterior	44
can	1429	rump	311	butts	42
rear	1107	fanny	206	fundament	11
arse	553	rear (a) end	199	tush	11
bum	510	tail end	122	derriere	7
butt	421				

9. More advanced queries

Most of the examples that we have seen involve single-word queries. For example, we have extracted synonyms, more specific words, less specific words, parts of whole, and whole for parts from WordNet, and have then seen the frequency of each of the words in these lists in the *BNC*. Recall, however, that in Sections 5 and 6, we used the WordNet information as part of a phrase. For example, we searched for [=wicked] [nn*] (all synonyms of [wicked] followed by a noun) or [=large] [=amount] (all synonyms of [large] followed by all synonyms of [amount]).

The ability to embed WordNet lists into phrase queries can be carried out for hypernyms, hyponyms, meronyms, and holonyms as well. For example, all of the following queries are possible:

Table 18: More advanced phrase-level queries.

<i>QUERY SYNTAX</i>	<i>MEANING</i>	<i>EXAMPLES</i>
[av0] [=hot] [<food]	adverb + synonym of [hot] + hyponym of [food]	really hot pizza incredibly spicy chicken
<eat] the [<food]	Hyponym of [eat] + the + hyponym of [food]	devour the steak munch the chips
br*k* my/his/her [@body]	forms of <i>break</i> + [my or his or her] + part meronyms of [body]	broke my nose breaks his ankle

One final extension of the query syntax is the ability for users to create “customized” or “user-defined” lists of words, which they can then re-use countless times in subsequent queries. These are already incorporated into the 100 million word Corpus del Español that I created in 2001 (see <http://www.corpusdelespanol.org>), and are a new feature in the BNC/VIEW interface. The basic idea is that users can create any number of lists of words that are morphologically, syntactically, or semantically related. They enter this list of words via the web-based interface, and can then re-use the list as part of subsequent queries – ten minutes or ten months later.

For example, if the user wants to focus on British English, s/he could modify the list of [body parts] from the American-based WordNet, and store this list of 80-100 words as his or her own list. Likewise, the user might wish to save just those synonyms of [beat] that refer to music, and then re-use this list in subsequent queries. Finally, in cases where WordNet is somewhat weak in terms of semantic fields, the user might want to create his or her own categories from scratch, such as [negative emotions] or [academic life].

Once created, these customized lists can then be easily incorporated into the standard query syntax. For example, suppose that [Jane.Smith] has created a list called [emotions] with 40-50 words like [happy, sad, excited, relieved], or that a user called [LingProf] has created a list of 80-100 computer-related terms like [mouse, CPU, laptop, web, spreadsheet]. The user then includes a reference to this list as part of the query string, e.g.

- (13) [av0] [Jane.Smith:emotions] *really happy, extremely angry*
 [vv*] the [LingProf:computers] *clicked the mouse, surfed the Web*

As we can see, the use of a relational database architecture means that we can add any number of levels of annotation or different modules, and then we simply create links between these to allow for very powerful queries.

10. More advanced queries involving register variation

Because the *WordNet* and *BNC* databases are in relational database form, they can easily be joined together with other databases in a way that would otherwise be quite impossible. For example, at <http://view.byu.edu/>, it is also possible to examine register variation in the *BNC*, with greater ease than with perhaps any other interface.

At the most basic level, one can find the frequency of a given word in all 70 registers of the *BNC* (e.g. for *kick* [v]), and sort the results by the relative frequency in these different registers. More advanced queries allow one to find which words occur with greater or lesser frequency in different registers. For example, in less than three seconds one could find:

- all adjectives that occur much more frequently in tabloid than broadsheet newspapers (e.g. *heartbroken*, *adults-only*, *hunky*, *smashing*);
- all verbs that occur more frequently in the [spoken:courtroom] section of the corpus than in other spoken registers (*harbouring*, *handcuffed*, *ascertained*, *disallowing*), or
- all nouns occurring more frequently in [w_ac_medicine] than in other academic texts (*colitis*, *ulcer*, *biopsy*, *gastrin*).

Because the ‘register’ databases contain the frequency of each word and *n*-gram in each of the 70 distinct registers of the *BNC*, it would be quite easy to combine this with the *WordNet* database. To follow up on some of the queries already presented in this paper, one could find, for example:

- which phrases with [synonym of *bad*] + [noun] occur more often in spoken than in written English (e.g. perhaps *big problem* or *bad luck*);
- more specific verbs expressing the concept [to walk] that are more common in fiction than in academic writing (e.g. *stagger*, *lurch*, and *trudge*);
- words relating to parts of the body that are more common in academic writing than in fiction or spoken English (e.g. perhaps *endocrine system*, *cervix*, or *thorax*).

One can easily imagine how such a corpus might be of value to both native speakers and to language learners. For example, language learners often encounter long lists of thematically-related vocabulary in a textbook, but there is often little indication of which words occur in more or less formal registers. As a result, they end up using a word like *buttocks* in very informal conversation, or conversely, a word like *fanny* or *bum* in more formal writing. The type of database that we have described would help them to easily check the correct register for a wide range of vocabulary items in different registers with one simple query.

11. Some limitations

In spite of the advantages of this approach, we should end the discussion by briefly considering two potential problems with the joint *BNC/WordNet* database. First, we should recognise that the database tends to overgenerate results. For example, [bottom] and [tail] show up as meronyms (parts of the whole) for [body], but probably only a small percentage of the occurrences of these words in the *BNC* refer to [the fleshy part of the human body that you sit on]. There is no way to automatically disambiguate word sense, as has been done manually, for example, with the 1 million word *SEMCOR Corpus* (cf. Landes et al. 1998), which has *WordNet* synset annotation for each word in the one million word *Brown Corpus*. However, by combining the one [body] “slot” in the search string with another word, the necessary disambiguation often occurs naturally. For example, if we add [broken] before [body]/[includes part], then nearly all of the hits do refer to the body: *broken leg, broken arm, broken neck*.

A second note concerns the content of *WordNet*. Because it is based primarily on American English and the *BNC* of course comes from the UK, there may at times be mismatches between the two. For example, the results from a search of the synonyms of [truck] fails to include [lorry], and [sweets] does not appear as a synonym of [candy]. In spite of the fact that *WordNet* is the most powerful semantically-based corpus in existence, there are still many gaps in its entries, especially to the degree that they do not reflect American English. This is why the ‘user-defined’ queries described in the previous section are so powerful. The end user can use *WordNet* as a starting point to create customised semantic fields and hierarchies, which can be modified in any number of ways for a particular end use.

12. Conclusion

We believe that the approach discussed here offers real advantages for a semantically-based investigation of the *BNC*. The relational database approach offer endless possibilities in terms of combining together frequency information for words and phrases, register variation, and semantic information. It is simply a matter of having the corpus creator write the correct SQL commands to perform the necessary SQL [JOIN]s from each table and database. One other advantage of this approach is that the queries are very fast. With the 100 million word *Corpus del Español* or the 100 million word joint *BNC/WordNet* database described here, even the most complex queries take no more than 3-4 seconds. In summary, our hope is that this database shows how – with the correct database architecture – one can begin to carry out advanced semantically-based, frequency-oriented investigations of large, diverse corpora.

References

- Aston, G. and Burnard L. (1998), *The BNC handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- Burnard, L. (2002), 'The BNC: where did we go wrong?', in: B. Kettemann (ed.) *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi. 51-70.
- Davies, M. (2003), 'Relational *n*-gram databases as a basis for unlimited annotation on very large corpora', in: K. Simov (ed.) *Proceedings from the workshop on shallow processing of large corpora*. Lancaster: Lancaster University. 23-33.
- Fellbaum, C. (ed.) (1998), *WordNet: an electronic lexical database*. Cambridge (MA): MIT Press.
- Landes S., C. Leacock and R. Teng (1998), 'Building semantic concordances', in: C. Fellbaum (ed.) *WordNet: an electronic lexical database*. Cambridge (MA): The MIT Press. 199-216.

This page intentionally left blank

Size matters – or *thus can meaningful structures be revealed in large corpora*

Solveig Granath

Karlstad University

Abstract

After sentence-initial thus, both S-V word order and inverted word order can be found. The standardised, million-word corpora of American and British English reveal four possible ways of ordering the main sentence constituents after initial thus: S-V, V-S, Aux-V-S, and Aux-S-V. However, larger corpora are needed to determine the reason for the variation. The British Guardian/Observer 1998-2002 (approximately 50 million words per year), used in the present investigation, shows some kind of inversion in 10-15% of the sentences with initial thus. A systematic comparison of examples with inverted and non-inverted word order demonstrates that this is neither a case of free variation nor formality of style, but rather a case of word order being used to signal a difference in the function of thus. Accordingly, S-V word order is used after resultative, summative, and appositive thus, whereas inversion is used when thus is a deictic proform.

1. Introduction

With the creation of the first computer-readable text corpora in the 1960s and 1970s, researchers were given a new tool for investigating different aspects of English. The first standardised million-word corpora to appear – the *Brown University Corpus (Brown)* and the *Lancaster-Oslo-Bergen Corpus (LOB)* – were huge by the standards of the time, and have proved very useful over the years for investigating common structures in English, not least because of the appearance of the updated versions compiled in Freiburg in the 1990s (*Frown* and *FLOB*). Nevertheless, we have come a long way since the early days of corpus linguistics, and even though the million-word corpora are still useful in many respects, it is important to have access to larger corpora, especially when it comes to investigating rarer structures. In this paper, I will give an example of a case when a million-word corpus is too small to provide the answer to a syntax problem.

2. The problem

What prompted this investigation was a sentence in the book *Alphabet to email* by Naomi Baron (2001). The topic of the section in which the sentence appeared was “The rise of RP”. At one point the author cites a quote from a tract written in

1798 to the effect that “public schools” can “correct the faults of provincial dialect”, and then comments (italics in this and other examples are mine):

- (1) Residential schools were up to the task. *Thus did* a particular variety of spoken English, long associated with the Court, London, and education at Oxford and Cambridge, *become* a nationwide standard, spoken by members of a limited social stratum, regardless of their original geographic provenance. (Baron 2001: 132)

What is interesting about the second sentence in this quote is that it does not have the regular Subject-Verb word order, but instead what I will here refer to as ‘partial inversion’, meaning that an auxiliary – in this case *did* – precedes the subject, and the main verb follows. Partial inversion is the term used in Biber et al. (1999: 911) to refer to this type of word order; other terms found are ‘Subject-operator inversion’ (Quirk et al. 1985: 1379) and ‘Subject-auxiliary inversion’ (Huddleston and Pullum 2002: 93). This is the word order that is normally used in yes-no questions in English, but we find the same word order also after a clause-initial word or phrase that negates or restricts the sentence, as in (2a), and after an initial intensifying *so* or *such*, as demonstrated in (2b):

- (2) a. *Only* this month *did* the scale of the crisis *become* clear.
(*The Guardian*, 2002)
- b. *So* central *did* this heavily embellished story *become* that Hitler Youth members were sworn in using the ‘Langemarck oath’.
(*The Guardian*, 2002)

However, the use of partial inversion after *thus* in (1) cannot be explained by either of these two cases. Our next step, then, is to see if grammars can supply any information on inversion after other clause-initial elements than the ones exemplified in (2). One clue may be provided by Quirk et al., who say that “[i]n older English [...], S-op inversion occurred without a preceding negative and some instances continue to be found in very formal style: “He invested the money – and *bitterly did* he *come* to regret it”. (Quirk et al. 1985: 1383; my emphasis). *Thus* is not mentioned in this context by Quirk et al.; however, Huddleston and Pullum (2002), in their discussion of inversion, include an example similar to (1), saying that “in relatively formal style inversion may occur following the preposing of a wide variety of elements”. The example they cite is *Thus had they parted the previous evening* (Huddleston and Pullum 2002: 96; their underscore).

Hence, according to these two grammars, the word order in (1) is essentially a formal variant of the normal, unmarked Subject-Verb word order. One problem with this explanation is that Baron’s book is written in a rather informal style, and there does not seem to be a reason for the writer to shift to a more formal style all of a sudden.

3. What corpora can reveal about word order after sentence-initial *thus*

Let us now see what corpora can reveal about the structure. First of all, an examination of all sentences with initial *thus* in the standard million-word corpora makes one thing immediately apparent, and that is that there are not two, but actually four alternative ways of ordering the subject and the verb after sentence-initial *thus*, namely Subject-Verb, Verb-Subject, Auxiliary-Subject-Verb, and Auxiliary-Verb-Subject, as shown in (3) (subject underlined):

- (3) a. *Thus a new pattern of days began* to develop, for Granny Albright did not die. (*Brown* K26 0250) (Subject-Verb word order)
- b. A devout and trusting people, they agreed to give up their home for a project that they were told was for the good of mankind. *Thus began their woeful saga* as nuclear nomads repeatedly relocated to other Pacific islands, where they have found only unhappiness. (*Frown* F24 34) (Verb-Subject word order)
- c. “Mine,” Nigel Lester had decided with a sigh of satisfaction, and then regarded the ring which he retrieved from his pocket. “At a price,” he concluded, with some bitterness. *Thus had they parted* the previous evening and now Diana was trailing up the gravelled drive to the hospital alone. (*LOB* P03 207) (Auxiliary-Subject-Verb word order)
- d. Clearly, then, this measure of strain is unsuitable for such large deformations, since 100 per cent compression certainly represents much more straining than does 100 per cent extension. *Thus can be realized the first result of having such large deformations*, namely the necessity for specifying some better measure of strain. (*FLOB* J71 126) (Auxiliary-Verb-Subject word order)

Here, (3a) has the unmarked Subject-Verb word order, and (3c) has partial inversion, as in (1) above. The examples in (3b) and (3d), by contrast, are cases of ‘full inversion’, i.e. the whole verb phrase precedes the subject.¹ Since these may be taken to represent two different cases of full inversion, one where the verb phrase consists of just one (main) verb, and one where it includes an auxiliary, they will here be presented as forming two different categories.

The next consideration is how common are these different types of inversion? Again, if we begin by examining all sentences containing initial *thus* in the four million-word corpora (Tables 1 and 2), it appears, first, that they are rather marginal – they occur less than 4 times out of a hundred – and second, that there has not been any great change in usage over time, since the figures for the two corpora from the 1960s do not differ greatly from the figures for the corpora from the 1990s.

Table 1: Word order after sentence-initial *thus* in the standard corpora.

<i>Corpus</i>	<i>SV</i>	<i>VS</i>	<i>Aux SV</i>	<i>Aux VS</i>	<i>Total</i>
<i>Brown</i> (AmE, 1961)	159	0	0	1	160
<i>LOB</i> (BrE, 1961)	115	3	2	0	120
<i>Frown</i> (AmE, 1992)	126	4	0	0	130
<i>FLOB</i> (BrE, 1991)	131	0	1	1	133

Table 2: Word order (in percentages) after sentence-initial *thus* in the standard corpora.

<i>Corpus</i>	<i>SV</i>	<i>VS</i>	<i>Aux SV</i>	<i>Aux VS</i>
<i>Brown</i> (AE, 1961)	99	0	0	1
<i>LOB</i> (BrE, 1961)	96	2.5	1.5	0
<i>Frown</i> (AmE, 1992)	97	3	0	0
<i>FLOB</i> (BrE, 1991)	98	0	1	1

Given, however, the small number of examples – inversion is used in no more than 12 sentences – it is not really possible to determine whether it is only the level of formality that decides which word order is used, as the grammars suggest, or whether there are other factors which influence the choice. This is where large corpora can be useful. In this particular case, I have used the two British newspapers *The Guardian/The Observer* on CD-ROM, from the years 1998 until 2002, each year consisting of approximately 50 million words. The results are interesting (Tables 3 and 4): not only can we note that some kind of inverted word order is used in 10-15 per cent of the sentences that begin with *thus*, we can also see that the figures for the different patterns remain fairly stable over the five years that the investigation spans.²

Table 3: Word order after sentence-initial *thus* in *The Guardian/The Observer* 1998-2002.

<i>Year</i>	<i>SV</i>	<i>VS</i>	<i>Aux SV</i>	<i>Aux VS</i>	<i>Total</i>
1998	672	50	63	9	794
1999	582	42	36	7	667
2000	764	43	47	5	859
2001	723	25	48	3	799
2002	685	35	29	10	759

Table 4: Word order (in percentages) after sentence-initial *thus* in *The Guardian/The Observer*.

Year	SV	VS	Aux SV	Aux VS
1998	85	6	8	1
1999	87	6.5	5.5	1
2000	89	5	5.5	0.5
2001	90.5	3	6	0.5
2002	90	5	4	1

As has been pointed out, some of the more recent grammars of English acknowledge that partial inversion may occasionally be found after non-negative opening adverb phrases. Full inversion, however, is not cited as an option.³ Nevertheless, statistics both from the standard million-word corpora and large British newspaper corpora show that not only partial inversion, but also full inversion is found after sentence-initial *thus*, and moreover, that the latter is used both when the verb phrase consists of a single verb, and when it includes an auxiliary. The lack of mention of these different alternatives in reference books makes this a problem that needs to be addressed, not least considering that full and partial inversion after *thus* is not just found in one or two odd examples, but occurs regularly in the large corpus used for this paper.

Given the frequency with which inversion is used after *thus*, the next question is, why is it used at all? Is it a marker of formality, as suggested in the grammars cited above? An examination of the material shows that it is sometimes used in fairly formal contexts, as in (4a) and (4b), in which the topics are religion and politics respectively; nevertheless, both full and partial inversion abound in texts dealing with sports (4c) and entertainment (4d), i.e. in contexts which must be characterised as being fairly informal.

- (4) a. “[...] And there he lets the hungry dwell, and they establish a city to live in; they sow fields, and plant vineyards, and get a fruitful yield.”
Thus states Psalm 107. (*The Observer*, 1999)
- b. Perrette gave de Gaulle a first-hand account of tanks in battle since the young captain had been a prisoner-of-war for the last 32 months of the first world war and knew little about them. *Thus did* Perrette initiate the future general into the subject. (*The Guardian*, 1999)
- c. We could have matched Woods’s effort out of the bunker too. He splashed out to six feet and missed the putt. *Thus ended* his dreams of a second consecutive Open championship. (*The Guardian*, 2001)
- d. This did not go unnoticed. Nor did the bandleader, when, a few weeks later, he fell off the stage. *Thus did* the 16-year-old get to lead the band – and to become the hotel chain’s musical director. (*The Guardian*, 1999)

As mentioned above, Quirk et al. refer to inversion being used “without a preceding negative” in older English (1985: 1383). This implies that it may be used in contexts that are archaic in style. Of the 229 sentences with full inversion in *The Guardian / The Observer* 1998-2002, a total of six are quotations from old texts. There are also modern texts that use old forms; hence, *saith* is used once, and the archaic *spake*, as in (5), is used 14 times, perhaps by analogy with the title of Nietzsche’s work *Thus Spake Zarathustra*.

- (5) Would he be going to a game on Saturday, Wilkinson asked. “I don’t know. I’ve got piles of things to catch up with,” he said. *Thus spake* a technical director and not the England coach. (*The Observer*, 1999)

In total, such archaisms make up less than ten per cent of the examples with inversion in the corpus.

To sum up, although examples can be found in which inversion is used in archaic as well as formal contexts, the majority of examples do not fall into these two categories, and we must therefore look for other reasons why inversion is used after *thus*. In the following sections, full and partial inversion will be examined separately in order to establish whether there is a reason for choosing one type of inversion rather than another.

4. Full inversion after sentence-initial *thus*

Altogether, there are 229 cases of full inversion in the newspaper corpus used for this study. Interestingly, the verbs used in these sentences fall very neatly into three semantic groups: report verbs, aspectuals (including the subgroup inchoative verbs), and the copula *be*.

Examining first the sentences containing verbs here referred to as ‘report verbs’, we find that by far the most frequent verb is *speak*, including the forms *speaks*, *spoke*, and *spake*; however, the range of verbs used (*argue*, *comment*, *go*, *observe*, *proclaim*, *read*, *run*, *say*, *scream*, *sing*, *soliloquise*, *state*, *think*, *write*) indicates that inversion after *thus* is used with report verbs in general.⁴ Most examples are of the VS kind, which means that the verb is in the simple present or past tense (6a, b), but occasionally, complex verb phrases with the word order Aux-V-S may be found (6c).

- (6) a. “The arrival of a baby often causes marital problems [...] in a minority of cases strains can lead to extra-marital affairs, post-natal depression, and even violence.” *Thus says* the Government’s Green Paper on the family. (*The Guardian*, 1998)
- b. “Oh no! My pretend boyfriend is gay!” *Thus ran* the Chinese title translation of last year’s Hollywood jolly *My Best Friend’s Wedding*. (*The Guardian*, 1998)

- c. [...] the minister's decision, if he reaffirms it, will be sheer vindictiveness, if not proof that those student days in Santiago make him devoid of impartiality. *Thus will sing* the piety of Latin America's new discovered champion, Lord Lamont. (*The Guardian*, 1999)

The second group is made up of aspectuals, verbs that signal the initiation, continuation, or termination of an activity. Most common in the corpus are *begin* and *end*, though other semantically related verbs also occur. Hence, to denote initiation, we find *begin*, *come*, *commence*, *start*, and *open*; continuation is indicated by *continue*, *develop*, *ensue*, and *follow*, and the three verbs *end*, *depart*, and *go* signify termination. Here we have another group large and varied enough to suggest that it is their meaning rather than any kind of idiosyncratic behaviour that makes them prone to appear in these inverted sentence structures. Similar to the sentences where report verbs are used, the majority of these examples are in the simple present or past, and therefore have VS word order, as in (7a). Nevertheless, there is one example in the present perfect, namely (7b), which has Aux-V-S word order.

- (7) a. He ended up buying a baseball team called the Los Angeles Angels. *Thus began* a very sweet, but ultimately unfulfilled love story. (*The Guardian*, 2002)
- b. At the same time, tanks and troops are building up on the borders of other cities, including Nablus and Hebron. *Thus has begun* the gradual reoccupation of the Palestinian areas by the Israeli military [...]. (*The Guardian*, 2001)

A subgroup of aspectuals is made up of inchoative verbs, i.e. verbs signalling that something comes into existence. Because they are very often in the passive or in the present perfect, most of the corpus examples in this group are of the Aux-V-S kind (*was born*, *was created*, *was established*, *was initiated*, *is built*, *will be set in motion*, *has evolved*, *has spread*). *Come* may be extended by a prepositional phrase or a particle into the complex *come into being*, *come into existence*, *come about*. A few verbs in this group, such as *emerge* and *unfold*, are usually in the simple present or past tense. Some examples of inchoative verbs are presented in (8).

- (8) a. A couple of years later, Diller hired another young executive, Michael Eisner, from TV network ABC where he had commissioned the hit series *Happy Days*. *Thus was initiated* the most propitious partnership in modern Hollywood: [...]. (*The Guardian*, 2001)
- b. [...] *Thus will be set in motion* a chain of events which will lead to the first and, if Fifa is to be believed, the last World Cup to be hosted by two countries, in this case Japan and South Korea. (*The Guardian*, 1999)

Finally, a little less than 10 per cent of the examples with full inversion contain the copula *be*, as shown in (9).

- (9) a. To be a vampire slayer, by contrast, is to be a fast-handed destroyer of evil, who, by day, has to do her homework and worry about her make-up. *Thus is Buffy.* (*The Guardian*, 1999)
 b. This time, for dramatic effect, Campbell dutifully and suitably ummed and aahed and pondered before giving the spot-on answers to the questions he already knew. *Thus is the old trooper, ace of spin and maestro at manipulating an audience.* (*The Guardian*, 2002)

I will leave a discussion of the motivation for using inversion in these cases until we have looked also at some examples of when partial inversion is used after *thus*.

5. Partial inversion after sentence-initial *thus*

In contrast to the neat semantic groups that the verbs in the sentences with full inversion can be divided into, the verbs in the examples that contain partial inversion have no semantic common denominator. The examples in (10) and (11), for instance, contain the verbs *prove*, *reveal*, *widen/deepen*, and *avoid*. When it comes to tense, aspect and voice, simple present, as in (10a), and simple past predominate, together with past passive, an example of which is included in (10b).

- (10) a. Another trainee, ‘Spencer’, failed to return to base, having got hopelessly drunk and forgotten about his assignments. *Thus do the James Bonds of tomorrow prove themselves.* (*The Guardian*, 2001)
 b. He said that to argue with the need for reduction was to be negative. And that was unacceptable. We had to be positive. *Thus was the hand beneath the puppet revealed.* (*The Observer*, 2001)

Less common, but still to be found, is partial inversion in the progressive (11a) and in the present perfect (11b), as well as in future and modal constructions.

- (11) a. The earlier US distinction between Palestinian resistance and al-Qaida-style global terror appears to have been abandoned, as the secular core of the Palestinian national movement is absurdly lumped together with the Taliban. *Thus is the war on terror being widened and deepened – [...] – in ways that seem certain to rebound on its instigators with deadly ferocity in due course.* (*The Guardian*, 2001)
 b. As a departure at the top – however unintended it may have been – it was magnificent. *Thus has Perec, with a petulant flourish and a nifty*

piece of blame displacement, *avoided* humiliation on the track. (*The Guardian*, 2000)

With no restriction on the semantics of the verb used, nor on tense, voice or aspect, the final question that needs to be asked here is whether inversion after *thus* is only a less frequent variant of the S-(Aux-)V word order, as Huddleston and Pullum (2002: 96) suggest. Let us take a look at some parallel examples and see what conclusions we can draw.

6. Semantic motivation for using inversion after *thus*

Because this paper is based on a large corpus (almost 4,000 sentences, 452 of which contain some kind of inversion), it is possible to compare examples where the same main verb but different word order is used. Interestingly, such a comparison strongly suggests that word order is used to signal a difference in the function of sentence-initial *thus*. The clearest examples are those containing report verbs. Compare (12a) with (12b):

- (12) a. ‘Not since Thirtysomething has a series so divided the nation, with half the viewers enthralled, half aghast.’ *Thus says* America’s *Entertainment Weekly* of this award-winning comedy-drama, [...] (*The Guardian*, 1998)
- b. [...] it is clear [...] that Dr Reid is seeking to agree a line with Mr Rowley which falls short of a full and comprehensive account of the events of which they both have knowledge. *Thus*, at one point Dr Reid *says* to Mr Rowley: “You don’t have to tell any lies. Do you know what I mean?” (*The Guardian*, 2001)

In both these sentences, *thus* functions as a connecting device between the preceding text and what follows. The difference between the two examples is that in (12a), *thus* serves as an anaphor for the words which have just been uttered (‘that’s what’ X uttered), whereas with Subject-Verb word order (in (12b)), *thus* heads a sentence which elaborates on a statement that has just been made, in this case by providing additional information that will serve to clarify the previous claim. It is important to note that what is reported (directly or indirectly) precedes *thus* when inversion is used, but follows the report verb with Subject-Verb word order. In the next pair of examples, sentences where the verb is *begins* are contrasted.

- (13) a. Clare thinks that “he should have brought his will up to date”. Jack is amazed that he could make such a fundamental change to his principles which, he believes, “should not be a moveable feast”. *Thus begins* their first quarrel. (*The Guardian*, 2002)

- b. Since his first solo effort, *Heartbreaker*, former Whiskeytown leader Ryan Adams has swapped New York for Los Angeles. *Thus* Gold begins with the ennui of *New York New York* (not that one), and comes to rest, 16 songs later, on the lullaby *Goodnight Hollywood Boulevard*. (*The Guardian*, 2001)

Again, in these two sentences *thus* has different functions. With inversion in (13a), *thus* has an anaphoric function ('in this way'), whereas in (13b), as in (12b) where there is no inversion, *thus* has an appositive function, i.e. it introduces additional information that will back up the previous claim.

Finally, let us compare two examples, one of VS (14a), and one of SV (14b) word order where the subject is a personal pronoun and the verb a form of the copula *be*:

- (14) a. At the Guardian, Bill Tucker was deputy to the then chief engineer, Bob Elderfield, taking over as chief engineer in 1982. *Thus was he* in charge when hot metal still ruled at the Guardian. (*The Guardian*, 1999)
- b. Rightwing journalists loved to compare his battered, noble features (he broke his nose in a crash early in his career) to those of an ancient gladiator, or emperor. *Thus he was* often presented as a national archetype. (*The Guardian*, 2000)

My suggestion here is that *thus* in (14a) has approximately the meaning 'as such' (he was in charge; namely as chief engineer), whereas in (14b), *thus* sums up the previous utterance ('consequently').

Examples like the ones just presented could be multiplied. I have found no sources that indicate that word order is a tool which will help disambiguate the function of *thus* in clause-initial position; nevertheless, the four different uses of *thus* that Quirk et al. (1985: 487, 635, 1470) differentiate between serve as a good starting-point for pinpointing the different functions that *thus* may have in English. My claim here is that the function of sentence-initial *thus* is signalled by word order as follows:

- Summative *thus* (= 'therefore, to conclude, to sum up'): Subject-Verb word order
- Resultative *thus* (= 'consequently, as a result'): Subject-Verb word order⁵
- Appositive *thus* (used to go from the general to the particular): Subject-Verb word order
- *Thus* used as a deictic proform: Inversion (full or partial)

What my examples show quite clearly, I think, is that unmarked word order is used in the first three cases (summing up, stating the result, giving more specific information to back up a claim that has just been made), whereas when *thus* is

used in deixis, either full or partial inversion is used. Which type of inversion is used depends to a large extent on the semantics of the main verb, so that full inversion is preferred if the main verb is a report verb, an aspectual, or the copula *be*; with other verbs, partial inversion is selected.

7. Evidence from the standard million-word corpora

The next question is how far the results presented above can be generalised. Albeit the *Guardian/Observer Corpus* is very large (about 250 million words in total), the English used may reflect the house style of the two newspapers rather than general usage. Therefore, to cross-check the results, let us return to the twelve occurrences found in *Brown*, *LOB*, *Frown* and *FLOB* and see how they tally with the findings from the large corpus.

Out of the twelve instances of inversion after *thus*, as many as nine have full inversion, and of these nine, seven are of the VS type; the remaining two have Aux-V-S word order. Apart from the copula *be*, all semantic groups presented in Section 4 above are represented: there are two report verbs, both of which are used in the simple past; the main verbs in the remaining seven examples are all aspectuals. As was the case in the newspaper corpus, the examples that have Aux-V-S word order both belong to the subgroup ‘inchoative verbs’.

As in the newspaper corpus, the report verbs follow immediately after a quote. It is not so surprising that one of the sentences from the smaller corpora contains the verb *speak* (*Thus spoke my revered, and now, alas, dead, friend /LOB F26 14/*), which was by far the most common verb in this group in the larger corpus. The second report verb is more interesting, since it is the function rather than the basic meaning of the verb that makes it such, as shown in (15):

- (15) The room – with its barre, mirrors and American Ballet Theatre posters from around the globe – is alive with dancers’ energy. They pose with an innate sense of performance, and both the photographer and the camera love them. “You don’t need me in this picture. I’m a third wheel.” “No, no. Come back over here. We need you in this picture.” “How does this look?” “Great. Perfect.” “Move your head this way.” “Like this?” “Great. Perfect. You guys are wonderful. Hold it.” “Are you sure you need me in this picture?” *Thus cavorted* Karena Brock-Carlyle, her husband John Carlyle and Gaye Baxley Manhattan, who last month officially became artistic directors of Savannah’s Ballet South community dance troupe. (*LOB A41 46*)

Normally, *cavort* would hardly be considered a report verb, but in this context, where the dancers prance around while chatting with the photographers, an action verb rather than a straightforward verb of saying is used (see e.g. Biber et al.

1999: 196). This is similar to the way verbs like *smile* and *grin* are used in combination with direct quotes.

Of the aspectuals, the ones that were the most common in the newspaper corpus are also represented in the four general corpora. There are three instances of *thus began* and one of *thus ended*. In addition, *Frown* also includes an occurrence of *go* used as an aspectual:

- (16) Allen, who had wanted to publish *Empty Mirror* for the better part of a decade, and who already had an introduction for the volume written by William Carlos Williams, could not have been happier with this turn of events and he went right to work on assembling the collection. *Thus went* one of the most active periods in Ginsberg's publishing history, [...] (*Frown* G60 25)

Finally, the two examples containing inchoative verbs are both of the form Aux-V-S, similar to the patterns found in the large corpus. One of them was quoted in (3d); the second is from *Brown*:

- (17) When he showed this model as his 'solution' as to how the Howe sewing machine operated, he was told he was 'wrong', and discovered to his amazement that the Howe Machine, which was unknown to him in detail, used two threads while the one that he had perfected used only one. *Thus was invented* the single thread sewing machine, [...] (*Brown* H26 0730)

The remaining three examples, all of which have partial inversion, contain verbs from other semantic groups. One example is quoted above in (3c); the others are *Thus did Bruto join the shades of Ponce* (*LOB* F25 99) and *Thus do the 'spirits of great events stride on before them'*. As can be seen, the examples in which inversion is used after *thus* in the general corpora very nicely corroborate the findings based on the larger corpus.

8. Topics for further research

The results of the present investigation give rise to a number of additional questions, some of which will be briefly touched upon here. One question is whether there is any variation between the different types of inversion. There are indications that this is the case. For instance, it appears that with inchoative verbs in particular, either full or partial inversion may be used, with little or no change in meaning:

- (18) a. The first business model for advertising on the net was the 'eyeballs' one, based on the notion that if you could attract millions of surfers to a site, it ought to be possible to sell advertising space on its pages.

Thus was born the banner ad – that block of 468 by 60 pixels which now infests the commercial web. (*The Observer*, 2001)

- b. Here was an activity in which everyone could participate, from City analyst to inner-city drop-out. *Thus was* the UK's newest bank *born*: a national data base in which people 'bank' the time they have available. (*The Guardian*, 2001)

Another construction where there appears to be variation is in cleft sentences. Hence, in cleft sentences with a preposed deictic *thus*, Subject-Verb word order, as in (19a), is more common than inversion; occasionally, however, Verb-Subject word order is found, as shown in (19b).

- (19) a. Professor Clyde Chitty writes: I first met Caroline Benn in 1966. I was in London, just down from Leicester University and keen to get involved with the Comprehensive Schools Committee. *Thus it was that* I joined the CSC, then operating from the family home in Holland Park Avenue. (*The Guardian*, 2000)
- b. This demonstration of Frank's interpretative and observational powers led, in 1940, to an invitation from Jones to share the leadership of British intelligence for the duration of the war. *Thus was it that* Frank's observation of minute changes in the position of a shadow on an aerial photograph led to the first identification of Germany's large defensive rotating radar towers, [...] (*The Guardian*, 1998)

Whether the writer actually has a choice of word order in cases like the ones presented in (18) and (19), or whether there are factors that restrict the choice, will have to be left for further research. So will the question whether inversion to signal the meaning of *thus* represents an innovation in Late Modern English. One of the supreme authorities on the history of English syntax, Visser (1969), states that "formerly there were four types of declarative sentences in simultaneous use" (Visser 1969: 1522). His examples, cited here as (20), were made up, but the implication of his statement is that these four types of word order were in free variation:

- (20) a. After his death *his queen reigned*.
 b. After his death *his queen did reign*.
 c. After his death *reigned his queen*.
 d. After his death *did his queen reign*.

Visser cites numerous historical examples which show that after a sentence-initial "non-negative adverbial adjunct", both full and partial inversion were commonly found, and the author claims that eventually, partial inversion ousted full inversion (1969: 1523-1526). If Visser is right, and inversion was always an option after an introductory adverb phrase in Early Modern English, then using

word order to signal function must be a more recent introduction in English. However, the sentence with initial *thus* which Visser cites to show that partial inversion was used more freely in the 17th century than today is not really interpretable without context (his example from Milton 1644 is quoted in (21a)). With the context added, in (21b), it is clear that this is yet another example of deictic *thus*.

- (21) a. *Thus did Dion Prusaesus ... counsel the Rhodians.* (Visser 1969: 1525)
 b. Such honour was done in those days to men who professed the study of wisdom and eloquence, not only in their own country, but in other lands, that cities and seigniories heard them gladly, and with great respect, if they had aught in public to admonish the state. *Thus did Dion Prusaesus, a stranger and a private orator, counsel the Rhodians against a former edict; [...]* (Milton: *Aeropagitica*)

Another topic that must be left for further research is what positions deictic *thus* can take in the clause. In order to investigate this, all instances of *thus* in a corpus would have to be examined, which makes it a rather laborious undertaking (in the present investigation, only examples with capitalised *thus* were collected). However, a small pilot study shows that deictic *thus* can also be cataphoric, i.e. it can point to information that is yet to be supplied. Cataphoric *thus* can be found in clause-final position, immediately followed by a colon; what follows identifies and explains that for which *thus* acts as a cataphor. In the majority of the sentences retrieved, the main verb is a report verb and *thus* points to a following quote, as in (22a). Occasionally, other verbs are found (see (22b)). Both (22a) and (22b) are clearly cases of deictic *thus*, however.

- (22) a. Time magazine *reported* the fire *thus*: “While firemen restrained the nearly blind British author from running into the fire, Huxley wept like a child”. (*The Guardian*, 2002)
 b. The chatroom *works thus*: your live image is at the bottom right of the screen, the other live images fill most of the rest of the screen. At the bottom there is space to type in words (if participants do not want to use the phone). (*The Guardian*, 2002)

Yet another aspect of the problem is whether inversion is only used to signal the meaning of *thus*, or whether the same applies for other deictic words. A preliminary study indicates that the latter is indeed the case; inversion is used also after deictic *there*, *then* and *so*, as witness the examples in (23):

- (23) a. John Rodda [...], [...] reported the cycling from Herne Hill stadium. *There began* what he calls his love affair with the Olympics. (*The Guardian*, 2000)

- b. The organisation's illustrious history looked in danger in the 1970s [...]. *Then began* a programme of expansion and refurbishment which helped create an organisation wealthy enough to shower pounds 34,000 on each of its members. (*The Guardian*, 2000)
- c. As a boy, he inherited a portfolio of shares from her, and *so began* an interest in the markets. (*The Guardian*, 1999)

To pursue these questions here would be going beyond the scope of the present study. Further research, however, may be able to supply fuller answers to the issues raised here.

8. Conclusion: what standard corpora can tell us

The title of this paper is 'Size matters', and I hope that I have succeeded in showing how large corpora can be very useful in identifying meaningful structures in English. In the much smaller million-word corpora, even though all the four word order patterns that are used after *thus* showed up, the number of examples was too small for any underlying rationale for the variation to be discerned. However, even in a study such as this one, the smaller general corpora are of use. The way the texts in these corpora are classified according to type makes it possible to check whether a certain feature is typical of a certain category of text. Table 5 below (Appendix) lists the text categories from which the examples in Table 1 were taken.

Even though the number of examples is very small, the information we get from the table is rather revealing, and actually supports my conclusion above that inversion after *thus* is not used primarily in formal writing. We see that the overwhelming majority of examples come from the categories F (Popular lore), and G (Belles lettres, biography, essays). One of the twelve examples is even from one of the fiction categories, P (Romance and love story), and one is from the more formal category J (Scientific writing). Only two out of twelve examples are from Newspaper texts (categories A and B).

Clearly, what we can learn from smaller corpora is not unimportant. Therefore, the title of my paper is best taken to mean 'matters having to do with size' rather than 'bigger is better', because in the end, combining evidence from large and small corpora can give us information that neither type of corpus could provide on its own.

Notes

- 1 'Full inversion' is the term used by Biber et al. (1999: 911). Quirk et al. (1985: 1379) call this type of word order 'subject inversion', and Huddleston and Pullum (2002: 97, 67) refer to it both as 'subject postposing' and 'subject-dependent inversion'.

- 2 In the statistics, I have excluded verbless sentences and sentences where initial *thus* serves as an intensifier of an adjective or an adverb. I have also omitted references to Nietzsche's *Thus Spake Zarathustra*.
- 3 Full inversion after preposed place adverbs, as in the example 'Here comes my brother' (Quirk et al. 1985: 1380) is best analysed as a separate case. One way in which this construction differs from inversion after *thus* is that if the subject is a pronoun, inversion is disallowed (cf. 'Here he comes' with '*Here comes he'). After *thus*, full inversion is used even when the subject is a pronoun (see example (14a) below).
- 4 Some of the verbs in this section have been culled from *The Guardian/The Observer* 1995 and 1996.
- 5 Quirk et al. (1985: 635) use the term 'resultive'.

References

- Baron, N. S. (2001), *Alphabet to email: how written English evolved and where it's heading*. London/New York: Routledge.
- Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman grammar of spoken and written English*. London: Longman.
- Huddleston, R. and G. K. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.
- Visser, F. Th. (1969), *An historical syntax of the English language. Volume 1. Syntactical units with two verbs*. Leiden: Brill.

AppendixTable 5: Inversion after sentence-initial *thus* according to text category.

<i>Text category</i>	<i>Brown</i>	<i>LOB</i>	<i>Frown</i>	<i>FLOB</i>
A: Press reportage	-	-	1	-
B: Press editorial	-	-	1	-
C: Press reviews	-	-	-	-
D: Religion	-	-	-	-
E: Skills, trades etc.	-	-	-	-
F: Popular lore	-	3	1	1
G: Belles letters	-	1	1	-
H: Miscellaneous	1	-	-	-
J: Scientific writing	-	-	-	1
K: General fiction	-	-	-	-
L: Mystery	-	-	-	-
M: Science fiction	-	-	-	-
N: Adventure	-	-	-	-
P: Romance	-	1	-	-
R: Humour	-	-	-	-

This page intentionally left blank

Inversion in modern written English: syntactic complexity, information status and the creative writer

Rolf Kreyer

University of Bonn

Abstract

Full-verb inversion in English has been the subject of a large number of studies in the recent and the less recent past. The present study tries to give a corpus-based account of this phenomenon within a discourse-functional framework. First, I will describe the influence of syntactic complexity and information status. However, I will argue that inversion should not merely be regarded as a means to ensure processability and flow of information. Instead, inversion should be understood as the result of a conscious choice on the part of the writer, who makes deliberate use of this rather rare syntactic phenomenon to serve certain superordinate functions, namely text structuring and what I call the immediate-observer effect, i.e. helping the reader to immerse into the discourse world. It will be shown that the distributions of weight and information status within inverted constructions can be understood as a result of these two superordinate functions.¹

1. Introduction

The term ‘inversion’ has been used as an umbrella term for a number of syntactically distinct constructions. It has been claimed, for instance, that “there exist on the order of 20 to 40 types of inverted sentences in English” (Green 1985: 117). Such a broad understanding of the term includes phenomena such as subject-auxiliary inversion and even existential-*there* constructions. In this paper the term is applied in a more restricted way. Following Dorgeloh (1997), I will understand inversion as “all those constructions in which the subject follows all of its verb phrase, i.e. a full (lexical) verb or copular be” (Dorgeloh 1997: 23). Cases of *there*-insertion as in (1), although they fit the definition, will not be considered in this paper, since it has been shown by various studies that *there*-insertion is a construction in its own right (see Birner 1996, Dorgeloh 1997, and Birner and Ward 1998).

(1) On the table there was a yellow ashtray.

Under the definition given above, then, inversions fall into one of the five formal categories of Prepositional-Phrase-, Participle-, Adjective-Phrase-, Adverb-Phrase- and Noun-Phrase-inversion, which are exemplified in examples (2) to (6) below (bold characters in this and other examples are mine):

- (2) **To his left** was a long, shadowy, cobbled passage running beside what looked like barred loose boxes. (ACV: 204)²
- (3) a. **Lying inside, wrapped in a clean woollen shawl**, was the smallest baby I had ever seen. (CK0: 2708)
 b. **Neatly screened by another low wall of lattice-work bricks** were a compost heap, bound in by wooden slats, and an empty metal incinerator. (GUF: 1803)
- (4) **Common to all the contributors, up to and including Nietzsche**, is their profound interest in the literature of ancient Greece. (H0N: 9)
- (5) **Ahead** was a short marble staircase, leading to what appeared to be a lecture-room on the next floor. (ANL: 2607)
- (6) **A rather unusual example** are the boias-frias in Brazil. (AN3: 608)

Inverted constructions as one kind of a large variety of word-order phenomena in English have received a considerable amount of attention over the last few decades, with approaches ranging from generative and transformational over functional to cognitive backgrounds. The present study tries to give a discourse-functional corpus-based account of the construction at issue. To this end, three ‘factors’ of influence will be investigated into: syntactic complexity, information status and the creative writer. The first factor, syntactic complexity, has been shown to be highly influential with regard to a number of word-order phenomena (see, for instance, Hawkins 1994, Arnold et al. 2000, and Wasow and Arnold 2003). Surprisingly, however, syntactic complexity has for the most part been ignored in previous studies on inversion. To my knowledge, the only exceptions are Hartvigson and Jakobsen (1974), some remarks on inversion on the background of the ‘Principle of Early Immediate Constituents’ in Hawkins (1994), and a very sketchy treatment in Birner (1996) and Chen (2003). Nevertheless, the question as to what extent syntactic complexity contributes to the reordering in inversion merits attention, the more so since it has been identified as a powerful factor with regard to other word order phenomena.

In contrast to syntactic complexity, the second factor, information status, has been discussed widely in the context of inverted constructions. Most studies in this area of research treat inversion as an information-packaging device which serves to maintain the order of given before new elements within an utterance (see, among others, Green 1980, Penhallurick 1984, and Rochement 1986). The most thorough account is Birner (1996). She posits the following pragmatic constraint on inversion: “the information represented by the preposed constituent must at least be as familiar within the discourse as is that represented by the postposed constituent” (Birner 1996: xi). As will be shown in the following, this constraint does not do justice to the data that underlie this study, and an alternative will be presented in this paper.

The third factor assumes a wider perspective: inversions are analysed with regard to their potential to achieve certain text-structuring and aesthetic effects. Some such ‘superordinate’ functions have been hinted at in previous studies. Bresnan (1994: 85), for instance, argues that locative inversion serves the “special discourse function of PRESENTATIONAL FOCUS [...], in which the referent of the inverted subject is introduced or reintroduced on the (part of the) scene referred to by the preposed locative” (see also Bolinger 1971 and 1977 and Hannay 1991). In a similar vein, Hetzron (1975: 348) posits a “presentative movement” which defers “elements meant to be remembered in the subsequent context [... to] a sentence-final position”. This function of “supporting the organisation of a discourse” is also described in Dorgeloh (1995: 224); in addition, she claims that inversions serve to establish a certain viewpoint. This viewpoint effect, for instance, may make the reader assume an “eyewitness perspective” (Dorgeloh 1995: 228), as in (7) below:

- (7) Gleaming in the winter sunshine, Bosnian Serb guns strung out along this ramshackle hilltop village point menacingly towards Sarajevo’s western suburbs, three days after the expiry of Nato’s ultimatum for heavy weaponry to be withdrawn or put under United Nations Control. [...] *Beside the Warriors lie piles of razor wire that should by now have enclosed the cannons and mortars.* (Dorgeloh 1995: 228)

In the present study, the two superordinate functions of text-structuring and viewpoint creation will be discussed in more detail. In particular I will show that the distributions of syntactic weight and information status within inverted constructions can be understood as a consequence of these two superordinate functions.

2. The corpus

The data that underlie the present study are taken from the two *BNC*-genres ‘written-academic’ and ‘prose-fiction’ (see Lee 2001). This choice is due to the fact that these two genres have been shown to exhibit significant functional and formal differences. As for former, Longacre (1983: 5) points out that the genres ‘narrative’ (similar to prose fiction) and ‘expository’ (similar to academic writing) are diametrically opposed with regard to the dimensions ‘+/- contingent succession’ and ‘+/- agent orientation’. Following Longacre, Smith (1985) shows that to a certain extent this opposition is mirrored in the exploitation of linguistic features. As for the formal differences, Biber (1989) (see also Biber and Finegan 1986) has shown that the academic and fictional texts of the *LOB*-corpus behave differently with regard to Biber’s statistically derived text-types: 75% of all academic texts fall into two text-type categories which do not contain any fictional text. Correspondingly, the vast majority of fictional texts in *LOB* pertains to the text-type ‘imaginative narrative’, which does not include any

academic-prose texts. Due to these functional and formal differences, we may assume that genre influences on inverted constructions (if any) will show up clearly in the two genres that underlie the present study. Of the two genres ‘academic-writing’ and ‘prose-fiction’, a subcorpus has been created that contains approximately 130,000 sentences of each genre, representing each of the texts within the respective genres.

The search for inversions in an unparsed corpus is – to say the least – not without problems. Usually, researchers use as a database those tokens that they come across during their everyday reading (what Esser 2002 calls the ‘butterfly method’) or they search manually in representative text samples or corpora. Both methods are far from ideal. While the first will not lead to a representative sample, the second is extremely tiresome and will most probably suffer from the limitations and unreliability of the human reader: they may miss certain tokens either because of tedium and lack of concentration or because, unconsciously, they have been primed to certain patterns which are related to a particular kind of inversion. An automated search seems to be far more preferable in this respect.

In the present study, a third way which combines the strengths of the human researcher and the computer was pursued. Although the *BNC* does not offer any direct information on inverted constructions, it provides the researcher with sufficient indirect information which turns out to be of use when searching for inversions. This indirect information is provided through the word-class tags in the *BNC*. In Kreyer (2006) I have shown that the vast amount of inverted constructions usually corresponds to a fairly low number of tag-sequences. There is no space here to go into the minutiae of the procedure, but let us consider one example as an illustration. A large number of inversions with fronted subject complement, for instance, is realised by the tag-pattern ‘AV0_AJ0_VB.’:³ the fronted subject complement itself is realised as an adjective phrase with an adverb-premodifier; the adjective phrase itself is immediately followed by a form of *be*. Below are given some examples of inversions that fit into this pattern:

- (8) **Also available is** a file for metal and a rasp for wood which we found very useful for shaping and smoothing these materials. (A16: 976)
- (9) **Equally personal is** Janet Smith’s The World Outside My Window, the first piece she has made since disbanding her company and the first in a long time that she has made to please herself. (A1D: 61)
- (10) **More curious is** the extreme Catholicism of their patronage. (A24: 132)
- (11) **Especially disadvantaged were** children, whose needs were dramatically highlighted by the Child Poverty Action Group and similar organisations. (A66: 444)

This may help illustrate how an unparsed corpus can automatically be scanned for syntactic phenomena. On the whole, 583 search strings were applied which yielded a total of 972 tokens.

3. Measuring syntactic complexity and information status

To describe the distribution of weight within inverted constructions, a simple word-counting algorithm was used that compares the lengths of the two mobile constituents by calculating the difference in number of words between the post- and the preposed constituent.

1. Count the number of words that form the postposed subject, $\#_P$
2. Count the number of words that form the fronted constituent, $\#_F$
3. Calculate the difference $D = \#_P - \#_F$

Consider the examples below as an illustration of how the algorithm works:

- (12) **Also available is a file for metal and a rasp for wood which we found very useful for shaping and smoothing these materials.** (A16: 976)

$\#_P$: 20 $\#_F$: 2 D: 18

A file for metal and a rasp for wood which we found very useful for shaping and smoothing these materials is also available.

- (13) **More curious is the extreme Catholicism of their patronage.**

(A24: 132)

$\#_P$: 6 $\#_F$: 2 D: 4

The extreme Catholicism of their patronage is more curious.

- (14) **Into that group came Leonard.** (A0P: 1116)

$\#_P$: 1 $\#_F$: 3 D: -2

Leonard came into that group.

The D-value obtained by the calculation can be interpreted in two ways: 1) the sign tells us whether the inverted construction is easier to process ($D > 0$; examples (12) and (13)) or vice versa ($D < 0$; example (14)), and 2) the absolute value following the sign allows us to order each token with regard to the extent to which it facilitates or impedes processing. The absolute value in (12), for instance, is much higher than that in (13), i.e. 18 as opposed to 4. In this case, (12) could be regarded as being more strongly motivated by reasons of processing than (13), since the inversion in (12) facilitates processing to a larger extent than the inversion in (13).

To describe the distribution of information status a fourfold taxonomy of 'givenness' has been applied. Following Kaltenböck (1998, 2000), I understand 'givenness' as 'retrievability from the preceding text', as shown under (15):

(15) **‘Givenness’ as retrievability**

Information is given if it is retrievable from the preceding discourse.
Information is new if it is not retrievable from the preceding discourse.

The taxonomy itself distinguishes the following four degrees of retrievability (‘>’ means ‘more retrievable than’):

- (16) directly retrievable > indirectly retrievable > irretrievable-anchored > irretrievable-unanchored.

Directly retrievable are those elements that exhibit overt cohesion with regard to the previous discourse, as in (17):

- (17) Contributions become partnership property unless the contract provides otherwise, and any stipulation providing for interest or other financial benefit for a partner in consideration of his contribution is **null and void**. *Also null and void is any stipulation releasing a partner from playing an active role in running the business.* (EEH: 717-718)

The quality of being null and void is mentioned in the first sentence, the fronted constituent in the inverted construction therefore is directly retrievable from the context.

I use the term ‘indirectly retrievable’ to refer to what elsewhere has been called ‘inferable’. In (18), the constituents in bold print are points on the same scale: the sentence immediately preceding the inversion evokes a scale of identifiability by mentioning the quality of being very easily identifiable. The proposed constituent denotes another point on this scale.

- (18) On the basis of the work done by other researchers on the structure of DNA strands, and my own identification of Briant Bodies Alpha and Beta, I was able to choose a conception that would produce certain clearly defined physical characteristics. Sex, hair, eye and skin color, physique. These were **very easily identifiable**. *Less easy to identify were personality traits which I believe are carried by Briant Bodies Gamma and Delta.* (AN8: 2099-2102)

Irretrievable elements may be of two kinds, namely anchored and unanchored. Irretrievable-anchored are those elements that refer to a concept, an entity, a location and so on that is not referred to in the previous discourse, but the element at issue contains another element which is part of the previous discourse, as in (19):

- (19) “These long stretches of **uplands** were once the old sheep walks,” Fen explained as their path began to climb. A few sheep still grazed there. Here and there in the close-cropped grass grew small blue harebells and

masses of fragrant pink thyme, and Robbie was surprised to see one or two sheep nibbling at the herb. “It’s the thyme that gives the local mutton its superb flavour,” Fen told her. “You must be sure to try it.” *Beyond the uplands lay a stretch of woodland*, and Robbie was glad of the shade. (HHA: 2797-2802)

The fronted constituent *beyond the uplands* denotes a location that has not been part of the previous discourse. However, this location is introduced in relation to another location that is part of the previous discourse. The fronted constituent therefore provides irretrievable information but is anchored in the previous discourse since this new location is introduced in relation to an already given location. If no such link or anchor exists the term ‘irretrievable-unanchored’ is used, as in example (20):

(20) The Awful Beast

Standing before her across the water was the most enormous beast Anabelle had ever seen. It had large powerful muscles covered by a sleek golden coat, a massive boxy head with two huge floppy ears on either side, and a broad muzzle with a grand black nose, and two big, quick brown eyes. (CFJ: 1-3)

Here the preposed constituent *standing before her across the water* does not contain any link to the previous text.

4. The distribution of syntactic complexity and information status in inverted constructions

Figure 1 shows how syntactic complexity is distributed in the 972 tokens.

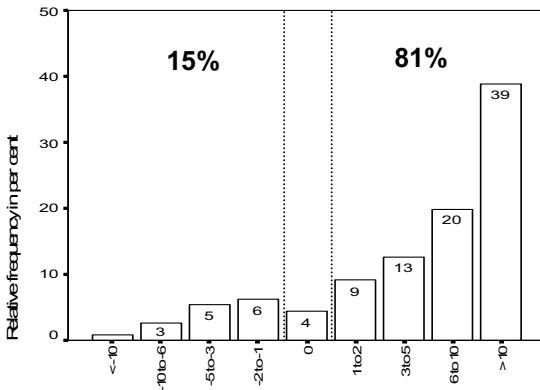


Figure 1: The distribution of complexity among inverted constructions.

As can be seen in Figure 1, 81% of all tokens have a positive D-value, i.e. in those cases the inverted construction is easier to process than its non-inverted counterpart. In 15% of all cases the canonical alternative is easier to process. In the remaining 4% the pre- and postposed constituent are of equal length and, accordingly, the variation of canonical and inverted construction does not lead to any changes in the distribution of weight. These numbers indicate a profound influence of syntactic complexity on inverted constructions.

Furthermore, in almost 40% of all cases, the postposed constituent is more than ten words longer than the preposed one; cases of extreme imbalance, as in example (21) below, are not at all uncommon in the corpus.

- (21) More seriously sacrilegious is surely Saint Pierre et le jongleur, Saint Peter and the jongleur, in which a jongleur's soul goes off to Hell with a number of other satirically identified characters jousting men, usurers, thieves, bishops, priests, monks, abbots, knights but presents itself, incongruously, as that of a relatively good character, anxious, for instance, to please its new infernal master (in a witty parody of the Orpheus story) by singing. (HXS: 202)

This uneven weight distribution among inverted constructions is also mirrored in the constituent lengths of pre- and postposed constituents. While with preposed constituents the maximum length is 24 words, it is 241 words for postposed constituents. Similarly, the mean constituent length is 5.06 words for preposed and 15.51 for postposed constituents, as is shown in Table 1.

Table 1: Minimum, maximum and mean length of pre- and postposed constituents.

<i>length of preposed constituent</i>			<i>length of postposed constituent</i>		
<i>min</i>	<i>max</i>	<i>mean</i>	<i>min</i>	<i>max</i>	<i>mean</i>
1	24	5.06	1	241	15.51

The data so far indicate a strong influence of syntactic complexity on inverted constructions, which could be described in terms of the principle of end-weight or a principle of light before heavy.

It might be expected that the distribution of information status in inverted constructions can be described in an analogous way, i.e. in inverted constructions given (or retrievable) elements precede new (or irretrievable) elements. This would nicely match the idea of Functional Sentence Perspective and the theme-before-rheme patterning. However, Table 2 shows that such a simple view does not do justice to the data.

Table 2: The distribution of information statuses across preposed and postposed constituents.

<i>information status</i>	<i>preposed constituent</i>		<i>postposed constituent</i>	
directly retrievable	30	3.1%	69	7.1%
indirectly retrievable	162	16.7%	78	8.0%
irretrievable anchored	679	69.9%	39	4.0%
irretrievable unanchored	101	10.4%	786	80.9%
<i>total</i>	972	100.1%	972	100%

Only 3.1% of all preposed constituents are in fact directly retrievable from the discourse and only 16.7% are indirectly retrievable. This means that over 80% of all preposed constituents are irretrievable, that is they denote entities, locations, qualities, etc. that are not part of the previous discourse. A similar proportion of retrievable and irretrievable elements is found with the postposed constituents, namely 15% vs. 85%. Obviously, then, inversions cannot be said to follow a simple given-before-new pattern: with both the pre- and the postposed constituents, the vast majority of items is situated at the ‘new’ end of the retrievability scale; the difference between the two, of course, is that postposed constituents are usually irretrievable-unanchored (80.9%), whereas the preposed ones usually contain an anchor to the previous discourse (69.9%) although they are ‘irretrievable’ in the strict sense of the term.

This seems to indicate that it is not the information of the pre- or postposed constituents *per se* that account for the felicity of inversion. Rather the crucial variable seems to be whether a constituent provides a link to the previous discourse or not. Seen from this perspective the fourfold taxonomy described above collapses into a dichotomy of ‘linked’ vs. ‘not linked’. The category ‘linked’ encompasses the first three information statuses, i.e. ‘directly retrievable’, ‘indirectly retrievable’, and ‘irretrievable-anchored’, while the category ‘not linked’ is identical with the category ‘irretrievable-unanchored’. Within this modified framework the distribution of information states across pre- and postposed constituents looks as depicted in Table 3:

Table 3: ‘Link’ vs. No ‘link’ across preposed and postposed constituents.

	<i>postposed constituent</i>					
<i>preposed constituent</i>	<i>no link</i>		<i>link</i>		<i>total</i>	
no link	78	8.0%	23	2.4%	101	10.4%
link	708	72.8%	163	16.7%	871	89.6%
<i>total</i>	786	80.8%	186	19.1%	972	100.0%

In the vast majority of all cases, i.e. 72.8%, the preposed constituent is linked to the previous discourse while the postposed constituent is not, with 8% of all tokens neither constituent is linked and with 16.7% both constituents are linked to

the previous discourse. Only in a few exceptions, namely 2.4%, is the postposed constituent linked to the discourse while the preposed constituent is not. These observations give rise to the following pragmatic constraint on inversion which I have formulated in analogy to and in lieu of Birner's constraint (1996):

(22) **Pragmatic constraint on inversions**

In those cases of inverted constructions where only one of the two mobile constituents provides a link to the previous discourse, this constituent is usually the preposed one.

In sum, the present data show that with regard to inversions, both 'syntactic complexity' and 'information status'/'linkedness' seem to be of influence. Inversion can therefore be regarded as being a consequence of the end-weight principle as well as of the pragmatic constraint formulated in (22). However, does this mean that inversions have to be interpreted as the result of the writer's compliance with two powerful word-order principles? In the following I will present an alternative motivation for inversions.

5. Text-structuring and observer-effect inversions

Inverted constructions are a rather rare and conspicuous phenomenon of written English. Since they deviate from the canonical word order of English, inversions are points of prominence within a text, and they will hit the eye of the reader. We may assume that a writer is aware of this particular quality of inversions and that he or she exploits the conspicuousness of this construction to achieve certain rhetorical or aesthetic effects: inversion may serve functions that are superordinate to merely ensuring processability and flow of information. In the present data, two functions of this kind could be identified: 1) inversions may serve text-structuring purposes by ensuring efficient integration of new information into the already existing text-structure, and by putting into final position those constituents that serve as the topic of the following discourse, and 2) inversions may be used to simulate 'natural' perception, thereby making the reader immerse into a narrative.

5.1 Text-structuring inversions

As has been pointed out above, the text-structuring function consists of two sub-functions: efficient integration of new information and topic managing. With regard to the first sub-function, consider (23), where the preposed constituent conveys irretrievable information but is anchored to the previous discourse. In such cases, the integration of the postposed constituent usually is a two-step process: first, the reader has to identify the site in the already existing discourse to which the postposed element should be attached, and second, it has to be specified in what way the postposed constituent should be attached to this site. This two-step process is realised by the preposed constituent that usually consists

of two components, namely a prepositional phrase (or another kind of postmodification) that names the linking site, and a further element that designates the linking relation.

- (23) 8 Some, as a result, have complained that regulations are “poorly enforced” and legislation is “ineffective” (Gunningham 1974: 56; also Freeman and Haveman 1973; Zwick and Benstock 1971), or that “a tradition of relatively weak enforcement prevails! (Bernstein 1955: 223). 9 ***Implicit in this stance*** is a conception that criminal law enforcement is properly a matter of compulsion, which leads, inexorably it seems, to a conclusion that regulation has failed. (FA1: 8-9)

In (23), a particular conception represented by the postposed constituent is to be introduced into the discourse. In a first step the preposed constituent tells the reader to what part of the previous discourse this conception should be related, namely the stance that is described in s-unit 8. Secondly, the adjective *implicit* tells the reader that this particular conception is a feature or an attribute of the stance mentioned in the prepositional phrase.

Example (23) illustrates a very important point: although the preposed constituent is irretrievable from the previous discourse, i.e. represents new information, it still may serve to integrate the postposed constituent into the discourse, since it contains an anchor to the previous discourse. The only necessary condition for the integration of information, then, is that the preposed constituent is in some way or other linked to the previous discourse, the exact information status is irrelevant.

The second sub-function is that of topic-management, i.e. the postponing of a constituent that serves as the topic of the following discourse. This is illustrated in (24).

- (24) 697 He drew in a deep icy breath, then straightened his shoulders, a habit he was forcing upon himself a lot of late, then made his way towards the kitchen, to be greeted by Mary with, “He’s gone then?” 698 “Yes, Mary, he’s gone.” 699 That was all he said, and he surprised her some. 700 what by walking quickly up the kitchen and into the hall. 701 He would generally stop and have a word or two or listen to her. 702 He was a good listener; he was about the only one in the house that was these days. 703 Everybody seemed to be in a rush. 704 It was the war, she supposed. 705 She got a bit lonely when she was on by herself at nights. 706 Things were changing in the house – you could feel it – and there was trouble brewing. 707 She had only to look at Mrs Jebeau’s face to see it; in fact, she could smell trouble in that direction. 708 She was a funny woman, was Mrs Jebeau, nervy; what they called neurotic, she would think. 709 “Yes” – she nodded to herself – “that was the word, neurotic, which accounted for her nerves and her funny temper too.” 710 *Going through Joe’s mind as he mounted the stairs were thoughts which were very similar, except that he*

expressed his in a slightly different way. 711 His mother, he knew, was in for another of her bouts, and he would have to bear the brunt of it. 712 He should be used to them by now because they had become a frequent occurrence during the past few months, particularly since Martin had been bringing Miss Crosbie to the house. 713 He had to think of her as Miss Crosbie so he wouldn't again make the mistake that had aroused his mother's anger when he had spoken of her as Marion. 714 As if he were a child of five she had reprimanded him, saying, "Don't be so personal; she is Miss Crosbie. 715 And don't address her by any other name." (CFY: 679-715)

The piece of text given in (24) is about the thoughts of two characters, Mary and Joe. S-units 699 to 709 portray an inner monologue of Mary. The inverted construction serves to shift the reader's attention away from this topic to another topic, which is denoted in the postposed constituent, namely the very similar thoughts of Joe. These thoughts then govern the text for the next five s-units.

5.2 Immediate-observer-effect inversions

In addition to the text-structuring function discussed above, inversions may also be used to mirror natural perception and thereby create the illusion on the part of the reader to be an immediate observer of a particular scene or event. This effect of inversions is due to the fact that inverted constructions are a useful means of mirroring natural perception. What does 'natural' mean in this context?

The process of looking at a particular object or at the layout of a room usually is a 'smooth' one: if for instance a man enters a room, he may find that there is a table in the middle of the room, he may then focus on the table and recognise a plate and a knife and a fork. Eventually, he may focus on the plate and find a steak there, and so on. Perception, then, is usually an activity that proceeds from one particular item to another item in its vicinity. Furthermore, perception may involve successive focusing: usually we perceive the bigger object first and then focus on particular parts of this object (see Kreyer forthcoming) for a more detailed discussion of this aspect).

These assumptions seem to be corroborated by an experiment by Ehrich and Koster (1983): they asked subjects to describe the layout of miniature living rooms and found a quite regular word-order pattern in the descriptions, namely: "reference to location almost always preceded reference to items being located [...]" (Ehrich and Koster 1983: 184). Obviously, inversions can be regarded as one possible instantiation of this pattern: first the writer establishes a particular location through, say, a prepositional phrase and then introduces an entity with regard to this location. This order of elements may be exploited for two different effects:

- it will be easier for the reader to imagine an object or a scenery if the description follows a natural order; the reader will be enabled to "construct

- a vivid representation of a perceptually inaccessible reference situation” (Drubig 1988: 88);
- in fiction, this effect will create the impression on the side of the reader to be part of the narrative and to experience objects, sceneries, events, etc. immediately through the eyes of a character and not through the description of the narrator.

Dorgeloh (1995: 228) uses the terms ‘eyewitness perspective’ and ‘camera movement effect’ (1997: 110). In both cases, the reader will usually have the impression of unmediated perception. I will use the term ‘immediate-observer effect’ to denote this particular potential of inverted constructions. Consider (25):

- (25) Ludo is conscientious. He bends closely to his work. He unscrews the plate and removes it from the door. *Behind the plate is a chiselled cavity. Inside the cavity is a polythene bag. Inside the bag are several smaller bags. Inside each of them is a single ounce of heroin.* (J13: 3398-3404)

Here, the succession of the four inversions creates a strong impression of experiencing the whole process of discovering the heroin; the reader participates into this process by recreating each of the steps one after another in their ‘natural’ order.

5.3 Syntactic complexity, information status and the creative writer

So far, a number of *prima vista* unrelated findings have been presented. The question that remains to be answered is that of the relationship between syntactic complexity, information status and the text-structuring and the immediate-observer-effect function of inversions. Are there any interdependencies and if so, how can they be explained?

To answer this question, in my view, we have to take into consideration that the kind of inversions discussed in this paper is, for the most part, a literal phenomenon. Therefore, inversions are tied to particular circumstances of production, which allow the writer to carefully choose the words and constructions he/she makes use of. Hence, it seems reasonable to assume that inversions are the result of a conscious choice on the part of the creative writer, who makes deliberate use of this conspicuous construction to achieve certain intended effects. Accordingly, the most promising starting-point for a discourse-functional account of inversion seems to be the discourse functions the writer supposedly had in mind, i.e. the text-structuring function and the immediate-observer-effect function. In the following, I will show that the distributions of syntactic complexity and information status that were found in the present data can be explained as deriving from the two superordinate discourse functions of inversion.

Let us take a closer look at the text-structuring function of inversions first. It has been shown that inversions serve two main purposes in text-structuring: 1) efficient integration of new information by preposing a constituent that contains a

link to the previous discourse, and 2) topic-managing by putting into final position those constituents that serve as a topic for the following discourse.

Under the assumption that the writer intends to enable efficient integration of information, it is of course no surprise that the majority of all text-structuring inversions contains a discourse-old link in the preposed constituent (89.6%; see Table 3). Far more interesting in this regard is the aspect of syntactic complexity, in particular the weight of the fronted constituent. It was shown that on average the preposed constituent was much shorter than the postposed one (5.06 as opposed to 15.51 words). This is not surprising: while the efficiency of integration depends on the front position of the linking element, it is also furthered if the linking element itself is not excessively long and complex: the shorter and lighter the linking element, the more efficiently can the link be established. The imbalance of syntactic complexity in favour of light preposed constituents can thus be regarded as derivative of the writer's intention to establish text-structural connections as early in the clause and as efficiently as possible.

As regards the topic-managing function of inversions, it is beyond reasonable doubt that the topic presented in the postposed constituent usually contains irretrievable information (84.9%; see Table 2): if it contained retrievable information, there was no need to integrate it into the discourse again. In addition, the postposed constituent usually does not contain a link to the previous discourse (80.8%; see Table 3). This can also be accounted for: if the postposed constituent contained a link to the previous discourse, there would have been no need for the writer to introduce it by an inverted construction. The subject itself would have maintained the link and the rest of the clause could already have elaborated on the subject.

Topic-managing also accounts for the heaviness of subjects in inverted constructions. As soon as a new topic has been introduced into the discourse, the writer will want to make the reader familiar with this new topic; he or she will want to establish this topic in the mind of the reader. As a consequence, the topic will be elaborated upon, and the most immediate way of elaboration is within the same phrase: the postposed constituent might become more complex.

So far it has been shown how the text-structuring use of inversions can explain the findings on syntactic complexity and information status. Similar explanations can also be provided for inversions in 'observer-effect' function. As said before, inverted constructions can be regarded as an iconic verbal representation of the process of natural perception by putting into initial position an item that either contains or represents a reference to an already perceived object. The postposed constituent is then introduced in relation to this first object. It goes without saying that the preposed constituent in observer-effect inversions will at least contain a retrievable link to the previous discourse, this link representing the previously focused item. In contrast, the postposed constituent will usually represent an irretrievable item: when a person surveys a scenery or an object, the perceiver will usually shift his/her attention from one point that he/she is currently focussed on to a new and so far unattended point in the visual field.

Accordingly, the postposed constituent in observer-effect inversions will usually represent irretrievable-unanchored information. Thus, the relative information status in observer-effect inversions can fully be accounted for by the writer's intention to mirror natural perception.

Similarly, the distribution of syntactic weight can be explained by recourse to the writer's intention to mirror a natural process. Imagine a person is surveying a scenery. At the moment he or she is focused on object A; after the scrutiny of this object, the surveyor may let his or her eyes roam and may find an object B in some location relative to object A. Being aware of the existence of this object, he or she may then focus on object B and take a closer look at its particular features. The process described above can be regarded as a succession of three distinct actions, namely 'scrutiny of object A', 'shift from object A to object B', and 'scrutiny of object B'. Inversions will be used to mirror the process in the middle, i.e. the shift from A to B. This stage would usually not contain any elaborations on A, since the scrutiny of this object has been concluded; neither would it contain any elaborations on object B, since this object will be scrutinised at a later stage. This is also reflected in the distribution of weight among the pre- and postposed constituent. The first is rather short since, at this stage of the entire three-step process, it has already had its due attention. The second is rather short too (at least in comparison to postposed constituents in text-structuring inversions), since it will be subject to closer scrutiny later.

6. Conclusion

The present study has tried to give a discourse-functional corpus-based account of inversion. The aspects under investigation included the influence of syntactic complexity and information status, and rhetoric and aesthetic functions of inversion. Although the phenomenon at issue can to some extent be explained as a consequence of the writer's compliance with word-order principles that ensure ease of processing and continuous flow of information, I have argued that a different point of view yields a more accurate description of the construction, namely the point of view of the writer as a creative user of the language. The writer makes deliberate use of inversions to achieve particular effects with regard to the reader, namely helping the reader to structure a text and to immerse in a narrative. It is these effects that the writer has in mind when choosing to use an inverted construction; the configurations of syntactic complexity and information status can be regarded as derivative of the writer's conscious exploitation of inversions for these ends.

A similar explanation may also hold for other word-order phenomena, such as fronting, *there*-insertion, *it*-extraposition, etc., which so far have been accounted by recourse to syntactic complexity or information status. Taking into consideration the creative writer and his or her intentions might suggest a different view on seemingly well-explained phenomena and call for a reassessment of the influence of basic word-order principles.

Notes

- 1 I would like to thank Jürgen Esser and Uta Schäpers for comments on earlier versions of this paper.
- 2 The code refers to the *BNC*: the three-letter code designates the *BNC*-document, the numbers following the colon refer to the s-units within the documents.
- 3 'AVO' stands for 'general adverb', 'AJ0' for 'general adjective' and 'VB.' for any verb form of BE. The pattern itself represents all s-units in the *BNC* with a sentence-initial adverb immediately followed by an adjective which again is immediately followed by a form of *be*.

References

- Arnold, J. E., T. Wasow, A. Losongco and R. Ginstrom (2000), 'Heaviness vs. newness. The effects of structural complexity and discourse status on constituent ordering', *Language*, 76: 28-55.
- Biber, D. (1989), 'A typology of English texts', *Linguistics*, 27: 3-43.
- Biber, D. and E. Finegan (1986), 'An initial typology of English text types', in: J. Aarts and W. Meijs (eds.) *Corpus linguistics II: new studies in the analysis and exploitation of computer corpora*. Amsterdam: Rodopi. 19-46.
- Birner, B. J. (1996), *The discourse function of inversion in English*. New York/London: Garland.
- Birner, B. J. and G. L. Ward (1998), *Information status and noncanonical word order in English*. Amsterdam: John Benjamins.
- Bolinger, D. (1971), 'A further note on the nominal in the progressive', *Linguistic inquiry*, 2: 584-586.
- Bolinger, D. (1977), *Meaning and form*. London/New York: Longman.
- Bresnan, J. (1994), 'Locative inversion and the architecture of universal grammar', *Language*, 70: 72-131.
- Chen, R. (2003), *English inversion: a ground-before-figure construction*. Berlin: Mouton de Gruyter.
- Dorgeloh, H. (1995), 'Viewpoint and the organisation of informative discourse. On the discourse function of full inversion in English', in: B. Warvik, S. K. Tanskanen and R. Hiltunen (eds.) *Organization in discourse: proceedings from the Turku conference*. Turku: University of Turku. 223-230.
- Dorgeloh, H. (1997), *Inversion in English. Form and function*. Amsterdam: John Benjamins.
- Drubig, H. B. (1988), 'On the discourse function of subject verb inversion', in: J. Klegraf and D. Nehls (eds.) *Essays on the English language and applied*

- linguistics on the occasion of Gerhard Nickel's 60th birthday*. Heidelberg: Groos. 83-95.
- Esser, J. (2002), 'Sampling and categorizing fronted constructions in the BNC', in: A. Fischer, G. Tottie and H. M. Lehmann (eds.) *Text types and corpora: studies in honour of Udo Fries*. Tübingen: Gunter Narr. 131-138.
- Ehrich, V. and C. Koster (1983), 'Discourse organization and sentence form: the structure of room descriptions in Dutch', *Discourse processes*, 6: 169-195.
- Green, G. M. (1980), 'Some wherefores of English inversions', *Language*, 56: 582-601.
- Green, G. M. (1985), 'The description of inversions in generalized phrase structure grammar', *Berkeley linguistics society*, 11: 117-146.
- Hannay, M. (1991), 'Pragmatic function assignment and word order variation in a functional grammar of English', *Journal of pragmatics*, 16: 131-155.
- Hartvigson, H. H. and L. K. Jakobsen (1974), *Inversion in present-day English*. Odense: Odense University Press.
- Hawkins, J. A. (1994), *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
- Hetzron, R. (1975), 'The presentative movement or why the ideal word order is V.S.O.P.', in: C. N. Li. Austin (ed.) *Word order and word order change*. Austin: University of Texas Press. 346-388.
- Kaltenböck, G. (1998), *Extraposition in English discourse. A corpus study*. PhD Thesis. Vienna: University of Vienna.
- Kaltenböck, G. (2000), 'It-extraposition and non-extraposition in English discourse', in: C. Mair and M. Hundt (eds.) *Corpus linguistics and linguistic theory*. Amsterdam: Rodopi. 157-175.
- Kreyer, R. (2006), *Inversion in modern written English: syntactic complexity, information status and the creative writer*. Tübingen: Gunter Narr.
- Kreyer, R. (forthc.), "'Observer effect", "eyewitness perspective", and "imaginary guided tour": What is so natural about the way inversions represent spatial relations?', *Cahiers de recherche en grammaire anglaise*.
- Lee, D. Y. W. (2001), 'Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle', *Language learning & technology*, 5: 37-72.
- Longacre, R. E. (1983), *The grammar of discourse*. New York: Plenum Press.
- Penhallurick, J. (1984), 'Full-verb inversion in English', *Australian journal of linguistics*, 4: 33-56.
- Rochemont, M. S. (1986), *Focus in generative grammar*. Amsterdam: John Benjamins.
- Smith, E. L. (1985), 'Text type and discourse framework', *Text*, 5: 229-247.
- Wasow, T. and J. Arnold (2003), 'Post-verbal constituent ordering in English', in: G. Rohdenburg and B. Mondorf (eds.) *Determinants of grammatical variation in English*. Berlin: Mouton de Gruyter. 119-154.

This page intentionally left blank

(3a) God have mercy on her benighted soul²

New fixed expressions (in the sense used in Moon 1998) can of course also appear at any time. One such came into English about two decades ago, when a hamburger chain called *Wendy's* began running a series of advertisements centered around the question

(4) Where's the beef?

implying that their competitors failed to have much more content than the seeded rolls you see on the outside. The theme was picked up by Walter Mondale in the 1984 elections, who in turn implied that his political opponent, Ronald Reagan, packaged and sold a similar product, all exterior and no content. Whether the *beef* idiom will be ephemeral or last remains, of course, to be seen.

In this paper, however we shall restrict our choice to a series of relatively well-known idioms, all of which are found in the *Collins COBUILD Dictionary of Idioms (CDI)* (2002) such as:

- (5) to kick the bucket
- (6) to die with your boots on
- (7) to bring coals to Newcastle

and examine the extent to which these prepackaged chunks of language can be internally expanded so as to link them into the discourse within which they are used. Thus, idiom (7) actually turns up in the *British National Corpus (BNC)* in the following form:

(7a) In their protest they sought to **restore some political coals to Newcastle**: the Declaration of Arbroath in 1320 committed the lay and ecclesiastical nobles of Scotland to support Robert Bruce, who stood against proud Edward's army, and sent him homeward to think again. (*BNC*: G1Y13)

Here, it is intuitively obvious both that coals are not normally *political*, and that the sentence nevertheless makes perfectly good sense.

2. Analytic framework

A first starting point may be found in the semantic analysis provided by Thomas Ernst in an article from 1980 called 'Grist for the linguistic mill: idioms and "extra" adjectives'. There, he distinguished among three types of adjectives that can modify idioms (Ernst 1980: 52-53):

- **external modification**, such as *Carter doesn't have an economic leg to stand on*. Here, the adjective refers to the surrounding (external) text.

- **internal modification**, such as *When will you get it through your small head that this isn't the way to do it?* The adjective modifies an element of the idiom, but does not change its reference. (Cf. example (3a), above.)
- **conjunction modification**, such as *Malvolio deserves almost everything he gets, but... there is that little stab of shame we feel at the end for having had such fun pulling his cross-gartered leg for so long.* Here, the underlined adjective reactivates the literal sense of the idiom, conjoining it to the figurative interpretation.

Ernst's analysis is concerned with the interpretation of the semantics of the idiomatic phrase, and he thus merely states in passing that the 'external' adjectives affect the entire idiom. The idiom is, however, more than merely modified: it is in fact hooked into or anchored in the discourse. It is this process that will be examined in the present article. Along the way, we shall see how common such anchorings are, and whether it merely is adjectives that can provide the hooks.

Moon notes that these anchoring adjectives "provide 'external modification', where the inserted adjective contextualizes the whole FEI [*Fixed Expression or Idiom*] in much the same way that a disjunct or sentence adverbial might" (Moon 1998: 174-175). Her corpus provides her with numerous examples, such as the following:

- (8) At our house, as ever, we are a little late **getting our Christmas act together**. (*OHPC: journalism*) (Moon 1998: 176)

which might easily be given the alternate form

- (8a) At our house, as ever, as Christmas approaches, we are a little late **getting our act together**.

This pattern is clearly more restricted than post-modification, which is particularly useful for allowing major expansion of an idiom, as in

- (9) (Jourdain) headed south and **bore the brunt of a ferocious storm which tore his mainsail and forced him to limp to the island of Madeira for a pitstop**. (*The Independent* 2000)

At the same time, inserting the 'hook' or 'anchor' into the middle of the idiom can make it seem less shopworn and give it more punch, as when the British soap character Zak Dingle notes:

- (10) Hey, nobody **rains on a Dingle parade!** (*Emmerdale Farm*)

as opposed to a potential alternative, post-modification:

(10a) Hey, nobody *rains on a parade* by the Dingle family!

Moon calls this type of extension ‘exploitations’, “the stylistic manipulation of the lexis (and semantics) of FEIs: perhaps to provide some sort of defamiliarization, and typically providing humour. [...] It is more marked and prevalent with metaphorical FEIs than any other type, since they contain the images which are most easily exploited.” (Moon 1998: 170).

In a relatively early corpus context, Ernst’s discussion of idiom modification was followed up by Nicolas (1995), who focused on adjectives inserted into the NP of V-NP idioms, introducing an ‘adverbial-based modifier classification’ with eight different potential (and five realised) types of such adjectives, basing his argument on introspection, complemented by data from the 1991 *Guardian* and 1987-1989 *Wall Street Journal*.³ His discussion is primarily a detailed analysis of the various types of modification and the structures of the NPs, together with their implications for the semantics of modification, with relatively little overall corpus data and no frequency data for the individual items.

Another useful distinction was made by Naciscione (2001), who works with ‘phraseological units’ in discourse, distinguishing at the discourse level between the ‘core use’, which is the simplex form used without further elaboration, and ‘instantial use’, which is when the idiom is used in a more elaborate form that “results in significant changes in its form and meaning determined by the context” (Naciscione 2001: 28). She notes that this is a ‘stylistic’ concept, and instantial uses are thus not copied verbatim by other writers. Harkening back to Geoffrey Chaucer once more, she states that “none of the 1164 instantial forms (in his works) became part of (the) phraseological stock of the English language” (Naciscione 2001: 31). Her discussion includes a number of examples of the instantial uses relevant to this paper.

3. The corpus material

To obtain a broad sample from various types of English, the following corpora were examined:⁴

- *BNC*
- *Los Angeles Times* (CD from 1990)
- *Broadcast News* (CD, radio transcripts from 1995)
- *The New York Times* (CD from 1996)
- *The Independent* (CD from 2002)

This provides material of about 300 million words, none of it overlapping. The various CDs have their own internal search engines, which made some searches difficult (they work best with idioms with highly distinctive components, such as *Hoist by/with one’s own petard*).⁵ The technical difficulties should not, however, have any major bearing on the results.

After an initial exploration of a number of idioms from the *CDI* in the *BNC*, 55 such expressions were then selected and searched for (Table 1):

Table 1: The idioms investigated.

<i>ace in the hole</i>	<i>new/fresh blood</i>	<i>fly the coop</i>
<i>hold all the aces</i>	<i>blot on the escutcheon</i>	<i>couch potato</i>
<i>acid test</i>	<i>soften the blow</i>	<i>paper over the cracks</i>
<i>balancing act</i>	<i>bone of contention</i>	<i>cut and dried</i>
<i>hidden agenda</i>	<i>die with your boots on</i>	<i>raw deal</i>
<i>whet your appetite</i>	<i>bear the brunt</i>	<i>just deserts</i>
<i>upset the applecart</i>	<i>drop in the bucket</i>	<i>a dime a dozen</i>
<i>out with the ark</i>	<i>bull in a china shop</i>	<i>back to the drawing board</i>
<i>babe in the wood(s)</i>	<i>bite the bullet</i>	<i>easier said than done</i>
<i>the whole ball of wax</i>	<i>carry the can</i>	<i>at a low ebb</i>
<i>baptism by/of fire</i>	<i>put the cart before the horse</i>	<i>catch her eye</i>
<i>bats in the belfry</i>	<i>cat and mouse</i>	<i>tempt fate/providence</i>
<i>amount to a hill of beans</i>	<i>skin a cat</i>	<i>at fever pitch</i>
<i>spill the beans</i>	<i>poisoned chalice</i>	<i>a finger in every pie</i>
<i>fit the bill</i>	<i>cheek by jowl</i>	<i>grey area</i>
<i>birds of a feather</i>	<i>chip on his shoulder</i>	<i>a drop in the ocean</i>
<i>champ at the bit</i>	<i>full circle</i>	<i>rain on your parade</i>
<i>blind as a bat</i>	<i>carry coals to Newcastle</i>	
<i>blood out of a stone/turnip</i>	<i>preach to the converted</i>	

In particular, it should be noted that, while the majority of these idioms have a pattern that easily allows the insertion of a modifier, others, such as *full circle*, *low ebb*, *cut and dried* or *couch potato*, are not as easily extended. This broad selection of different types of idiom patterns was deliberate, since it was likely to be of interest to see whether a variety of idioms could be anchored in the text in various fashions. Moreover, some of them can clearly be seen as more or less limited phrases, so that the citation form for e.g. *blood out of a stone/turnip* could also be considered to be *squeeze/get blood out of a stone/turnip*.

In the search, a fair amount of variation was permitted as regards form: if the idiom turned up via a search of the salient elements, it was normally included, even if it was rather different than the canonical form. An example:

- (11) More important is the fact that nobody can really tell whether Germany has a successful economy because of its voting system (though, between

you and me, this seems *as unlikely as a cart pulling a horse*) or whether it seems to have satisfactory constitutional arrangements because its economy has been working so well for so long. (*BNC*: AKR 888)

Since the corpora were not examined *in toto*, some of the more *recherché* examples may well have been missed.

4. Categorisation of the data

The primary distinction at stake here is whether the idiomatic expression has been modified such that it is anchored in the surrounding text or not. In its canonical form, an idiomatic expression is self-contained and sufficient unto itself, primarily serving as a means of summarising or commenting upon a situation, rather than providing more concrete information about it. It is this prototypically independent status that makes any overt linking to the text so noticeable.

The first category will thus be the ‘simplex’ or canonical form, with no extension beyond the basic syntactic shifts between positive and negative, active or passive and so forth. Even this category is not completely straightforward, however, since a number of idioms contain e.g. possessives, so that *to catch X’s eye* will normally include a component determined by the discourse:

- (12) Although it was the dazzling play in the backs that *caught the eye* that day, Britt was perhaps the most influential player on the field. (*Ind* 2002)
- (13) Colin Healey, 22, who *caught McCarthy’s eye* after joining Coventry City on loan from Celtic, [...] will be particularly disappointed. (*Ind* 2002)
- (14) It is not, you would think, the ideal way to *catch the eye of the great Britain coach David Waite* for this autumn’s series against New Zealand. (*Ind* 2002)

Here, the examples move from the abstract form *the eye* to a specific *s*-genitive form, to the (potentially much longer) postpositioned *of*-genitive. This anchoring in the text, however, is a relatively straightforward reference-tracking device, and no more surprising than Ernst’s type 2 adjective expansion, as in:

- (15) “Yet we cannot get visas and we are *such a small drop in the ocean* compared with the businessman.” (*BNC*: HJG 1067)
- (16) We munched our way through an average 18 pasta meals per head last year, *a mere drop in the ocean* compared to the Italians, who managed to swallow a massive 300 meals each. (*BNC*: K37 23)

Here, we can also add time adverbials, intensifiers and quantifiers, whether modifying the noun or the verb. Thus,

- (17) But as far as I can tell, people are still **chomping at the bit** to build here. (LAT 1990)

What is more problematic is to decide how to handle some of the various types of additional information that appear in connection with these idioms. Consider:

- (18) Having been at the end of a queue of Germans and Austrians all **chomping at the bit** to go heliskiing, it was our turn. (BNC: G2W 471)
- (19) It's countdown time to Croke Park and Derry footballers are **chomping at the bit** in their bid to make sure that "Sam" remains in Ulster for yet another year. (BNC: K32 1994)
- (20) US carriers and Lufthansa, among others, are **chomping at the bit** for new Berlin and East German air routes now that the air corridors, carved out in 1948, are becoming obsolete. (BNC: A94 196)
- (21) These dates have now become a **bone of contention** in Hong Kong as rugby administrators in the British territory have belatedly realized... (BNC: CB2 1750)
- (22) Under the old two-leg system, Spartak would not have **had a leg to stand on** after their 5-0 drubbing in the first meeting. (Ind 2002)
- (23) That achievement **comes full circle** as young people who have never been on stage before perform in front of thousands at the Kennedy Center. (BN 1995)

It seems reasonable to assume that (18)-(20) modify the idiom, and are merely forms of 'modification', whereas *in Hong Kong* in (21) is clearly peripheral to the idiom, although not to the discussion, and can easily be moved, e.g. to the front of the sentence:

- (21a) In Hong Kong, these dates have now become a **bone of contention** as rugby administrators in the British territory have belatedly realized...

Similar reasoning applies for the temporal clause in (22), and arguably even for the *as* clause in (23). The latter, however, is fairly close to the following example, where the second clause exemplifies the first:

- (24) My father taught me *the acid test* of a gentleman: respect for those who can be of no possible service to him. (*Ind* 2002)
- (25) Critics of Gorbachev ask why he does not *bite the bullet* and move directly to a market economy – as Poland has. (*LAT* 1990)

These seem to be tightly enough linked to the idiom to be worth keeping track of as a separate group, marked by being independent clauses, usually connected via *and* or a colon. We may call them ‘exemplifications’, as they amplify the idiom by offering a concrete example of its meaning.

Another group that should be separated out concerns cases where the idiom is in fact the name of a book, play, movie, TV show, song or the like. Names are relatively unproblematic in this context, although worth noting, in that they help keep the idiom alive and in circulation.

The result is a five-category system: simplex, modified (whether via premodification or postmodification), names, exemplifications and the subject of our primary interest, ‘anchorings’. (This latter category is what Nicolas (1995: 242) designates as the ‘viewpoint’ category.) These categories are here treated as mutually exclusive, although anchorings clearly are a subgroup of modification, distinguished only by their discourse semantics.

5. Results

The basic results are seen in Table 2.⁶

Table 2: Distributive percentages for the idioms investigated.

<i>Reference word</i>	<i>Canonical form</i>	<i>Tokens</i>	<i>Simplex forms %</i>	<i>Modifiers %</i>	<i>Anchorings %</i>	<i>Names %</i>	<i>Exemplifications %</i>
	Totals	7711	47.9	43.7	2.7	2.7	3.0
ace	ace in the hole	65	33.8	35.4	0.0	18.5	12.3
aces	hold all the aces	9	66.7	33.3	0.0	0.0	0.0
acid	acid test	118	30.5	53.4	2.5	8.5	5.1
act	balancing act	360	20.0	66.4	6.9	1.7	5.0
agenda	hidden agenda	297	48.1	35.	3.4	10.8	2.4
appetite	whet your appetite	258	22.1	75.2	2.3	0.0	0.4
applecart	upset the applecart	55	61.8	20.0	7.3	5.5	5.5
ark	out with the ark	5	80.0	20.0	0.0	0.0	0.0

babe	babe in the wood(s)	18	38.9	33.3	16.7	11.1	0.0
ball	whole ball of wax	13	84.6	7.7	7.7	0.0	0.0
baptism	baptism by/of fire	87	41.4	44.8	11.5	0.0	2.3
bats	bats in the belfry	9	88.9	11.1	0.0	0.0	0.0
beans	amount to a hill of beans	37	62.2	37.8	0.0	0.0	0.0
beans	spill the beans	100	47.0	47.0	3.0	2.0	1.0
bill	fit the bill	290	58.6	41.0	0.3	0.0	0.0
birds	birds of a feather	53	39.6	5.7	7.5	45.3	1.9
bit	champ at the bit	72	33.3	63.9	0.0	0.0	2.8
blind	blind as a bat	13	61.5	30.8	7.7	0.0	0.0
blood	blood out of a stone/turnip	24	87.5	8.3	4.2	0.0	0.0
blood	new/fresh blood	266	51.9	41.0	3.4	1.9	1.9
blot	blot on the escutcheon	5	20.0	20.0	40.0	0.0	20.0
blow	soften the blow	209	33.0	59.3	7.7	0.0	0.0
bone	bone of contention	133	25.6	69.9	3.0	0.8	0.8
boots	die with your boots on	20	65.0	5.0	0.0	25.0	5.0
brunt	bear the brunt	706	3.0	96.3	0.7	0.0	0.0
bucket	drop in the bucket	132	48.5	45.5	4.5	0.8	0.8
bull	bull in a china shop	26	80.8	11.5	7.7	0.0	0.0
bullet	bite the bullet	171	42.7	14.0	3.5	3.5	36.3
can	carry the can	63	50.8	41.3	3.2	1.6	3.2
cart	put the cart before the horse	85	80.0	10.6	4.7	0.0	4.7
cat	cat and mouse	214	67.3	27.6	3.3	1.4	0.5
cat	skin a cat	22	27.3	59.1	4.5	0.0	9.1
chalice	poisoned chalice	46	52.2	34.8	6.5	0.0	2.2
cheek	cheek by jowl	86	29.1	48.8	0.0	20.9	1.2
chip	chip on his shoulder	125	55.2	34.4	9.6	0.0	0.8
circle	full circle	408	58.8	29.2	1.7	2.5	7.8
coals	carry coals to Newcastle	14	64.3	7.1	7.1	0.0	21.4
converted	preach to the converted	69	69.6	18.8	0.0	0.0	11.6
coop	fly the coop	17	76.5	17.6	5.9	0.0	0.0
couch	couch potato	237	70.9	24.9	1.7	2.5	0.0
cracks	paper over the cracks	47	34.0	48.9	12.8	0.0	4.3
cut	cut and dried	112	53.6	44.6	0.0	1.8	0.0
deal	raw deal	152	53.9	35.5	2.0	5.9	2.6
deserts	just deserts	46	82.6	10.9	0.0	4.3	2.2

dime	a dime a dozen	45	97.8	2.2	0.0	0.0	0.0
drawing	back to the drawing board	234	60.7	22.2	2.1	0.0	15.0
easier	easier said than done	214	90.2	7.9	0.5	0.0	1.4
ebb	at a low ebb	195	39.0	59.0	0.5	0.0	1.5
eye	catch her eye	899	72.0	27.6	0.4	0.0	0.0
fate	tempt fate/providence	97	47.4	50.5	0.0	2.1	0.0
fever	at fever pitch	180	48.3	31.7	0.6	18.9	0.6
finger	a finger in every pie	24	62.5	25.0	8.3	0.0	4.2
grey	grey area	367	38.1	57.5	1.4	0.5	2.5
leg	have a leg to stand on	39	61.5	15.4	23.1	0.0	0.0
ocean	a drop in the ocean	89	43.8	50.6	4.5	1.1	0.0
parade	rain on your parade	34	35.3	23.5	17.6	23.5	0.0

Perhaps the most striking overall result is that the filling in the sandwich is so meagre, weighing in at only 2.7% anchorings, exactly the same level as for names. In consequence, removing the names will have relatively little effect on the anchorings, normally <0.2% for individual items, and 0.1% for the overall figures. As the relative figures for the other categories can nevertheless show considerably larger changes, Table 3 is also provided, giving the figures for the idioms, but with the names removed:

Table 3: Distributive percentages for the idioms investigated (names removed).

<i>Reference word</i>	<i>Canonical form</i>	<i>Tokens</i>	<i>Simplex forms %</i>	<i>Modifiers %</i>	<i>Anchorings %</i>	<i>Exemplifications %</i>
	Totals	7502	49.2	44.9	2.8	3.1
ace	ace in the hole	53	41.5	43.4	0.0	15.1
aces	hold all the aces	9	66.7	33.3	0.0	0.0
acid	acid test	108	33.3	58.3	2.8	5.6
act	balancing act	354	20.3	67.5	7.1	5.1
agenda	hidden agenda	265	54.0	39.6	3.8	2.6
appetite	whet your appetite	258	22.1	75.2	2.3	0.4
apprecart	upset the apprecart	52	65.4	21.2	7.7	5.8
ark	out with the ark	5	80.0	20.0	0.0	0.0
babe	babe in the wood(s)	16	43.8	37.5	18.8	0.0
ball	whole ball of wax	13	84.6	7.7	7.7	0.0
baptism	baptism by/of fire	87	41.4	44.8	11.5	2.3

bats	bats in the belfry	9	88.9	11.1	0.0	0.0
beans	amount to a hill of beans	37	62.2	37.8	0.0	0.0
beans	spill the beans	98	48.0	48.0	3.1	1.0
bill	fit the bill	290	58.6	41.0	0.3	0.0
birds	birds of a feather	29	72.4	10.3	13.8	3.4
bit	champ at the bit	72	33.3	63.9	0.0	2.8
blind	blind as a bat/&c	13	61.5	30.8	7.7	0.0
blood	blood out of a stone/turnip	24	87.5	8.3	4.2	0.0
blood	new/fresh blood	261	52.9	41.8	3.4	1.9
blot	blot on the escutcheon	5	20.0	20.0	40.0	20.0
blow	soften the blow	209	33.0	59.3	7.7	0.0
bone	bone of contention	132	25.8	70.5	3.0	0.8
boots	die with your boots on	15	86.7	6.7	0.0	6.7
brunt	bear the brunt	706	3.0	96.3	0.7	0.0
bucket	drop in the bucket	131	48.9	45.8	4.6	0.8
bull	bull in a china shop	26	80.8	11.5	7.7	0.0
bullet	bite the bullet	165	44.2	14.5	3.6	37.6
can	carry the can	62	51.6	41.9	3.2	3.2
cart	put the cart before the horse	85	80.0	10.6	4.7	4.7
cat	cat and mouse	211	68.2	28.0	3.3	0.5
cat	skin a cat	22	27.3	59.1	4.5	9.1
chalice	poisoned chalice	44	54.5	36.4	6.8	2.3
cheek	cheek by jowl	68	36.8	61.8	0.0	1.5
chip	chip on his shoulder	125	55.2	34.4	9.6	0.8
circle	full circle	398	60.3	29.9	1.8	8.0
coals	carry coals to Newcastle	14	64.3	7.1	7.1	21.4
converted	preach to the converted	69	69.6	18.8	0.0	11.6
coop	fly the coop	17	76.5	17.6	5.9	0.0
couch	couch potato	231	72.7	25.5	1.7	0.0
cracks	paper over the cracks	47	34.0	48.9	12.8	4.3
cut	cut and dried	110	54.5	45.5	0.0	0.0
deal	raw deal	143	57.3	37.8	2.1	2.8
deserts	just deserts	44	86.4	11.4	0.0	2.3
dime	a dime a dozen	45	97.8	2.2	0.0	0.0
drawing	back to the drawing board	234	60.7	22.2	2.1	15.0

easier	easier said than done	214	90.2	7.9	0.5	1.4
ebb	at a low ebb	195	39.0	59.0	0.5	1.5
eye	catch her eye	899	72.0	27.6	0.4	0.0
fate	tempt fate/providence	95	48.4	51.6	0.0	0.0
fever	at fever pitch	146	59.6	39.0	0.7	0.7
finger	a finger in every pie	24	62.5	25.0	8.3	4.2
grey	grey area	365	38.4	57.8	1.4	2.5
leg	have a leg to stand on	39	61.5	15.4	23.1	0.0
ocean	a drop in the ocean	88	44.3	51.1	4.5	0.0
parade	rain on your parade	26	46.2	30.8	23.1	0.0

Although the overall figures are quite low, a few idioms stand out for both number and percentage of anchorings. Here are the top 10, again with the names removed:⁷

Table 4: The 10 highest-frequency anchorings.

<i>Idiom</i>	<i>Nr</i>	<i>% anchorings</i>
have a leg to stand on	39	23.1
rain on your parade	34	23.1
babe in the wood(s)	18	18.8
birds of a feather	53	13.8
paper over the cracks	47	12.8
baptism by/of fire	87	11.5
chip on his shoulder	125	9.6
a finger in every pie	24	8.3
bull in a china shop	26	7.7
soften the blow	209	7.7

But even among these items, the most frequent amount to less than one in four instances, whereas a dozen idioms have (other forms of) modification in more than 50% of their occurrences. In addition, these frequencies suggest that some of the idioms are truly self-contained, hardly admitting of modification at all (16 are more than 2/3 simplex, with another 17 over 50% simplex), while others demand modification or exemplification. The highest frequencies are as follows:

Table 5: The 10 highest-frequency simplex idioms.

<i>Idiom</i>	<i>Nr</i>	<i>% simplex</i>
a dime a dozen	45	97.8
easier said than done	214	90.2
bats in the belfry	9	88.9
blood out of a stone/turnip	24	87.5
die with your boots on	15	86.7
just deserts	44	86.4
whole ball of wax	13	84.6
bull in a china shop	26	80.8
out with the ark	5	80.0
put the cart before the horse	85	80.0

Table 6: The 10 highest-frequency idioms with modifiers.

<i>Idiom</i>	<i>Nr</i>	<i>% modifiers</i>
bear the brunt	706	96.3
whet your appetite	258	75.2
bone of contention	132	70.5
balancing act	354	67.5
champ at the bit	72	63.9
cheek by jowl	68	61.8
soften the blow	209	59.3
skin a cat	22	59.1
at a low ebb	195	59.0
acid test	108	58.3

Table 7: The 10 highest-frequency idioms with exemplification.

<i>Idiom</i>	<i>Nr</i>	<i>% exemplification</i>
bite the bullet	165	37.6
carry coals to Newcastle	14	21.4
blot on the escutcheon	5	20.0
ace in the hole	53	15.1
back to the drawing board	234	15.0
preach to the converted	69	11.6
skin a cat	22	9.1
full circle	398	8.0
die with your boots on	15	6.7
upset the applecart	52	5.8

Not surprisingly, these lists are nearly mutually exclusive, with only 4 of 36 items occurring twice (although in different constellations): *bull in a china shop* (simplex, anchoring), *die with your boots on* (simplex, exemplification), *skin a cat* (modifier, exemplification), *soften the blow* (modifier, anchoring).⁸

Other factors must also be considered, as well. One structure that actively discourages further modification is when the idiom is inserted into an NP as a premodifier, as in

- (26) MILES O'BRIEN, Anchor: You don't have to run a marathon or pump up like Arnold Schwarzenegger to get a good workout, but you do have to leave your *couch potato* days behind. (*BN* 1995)

Here, the difficulties of layered processing would appear to be sufficient explanation for the difficulty of further modification, exemplification or anchoring.

Bite the bullet and *carry coals to Newcastle* are particularly prone to exemplification, i.e. having an additional clause instantiating what was done or must be done. And in a sense, they point the way forward to what is probably a more fruitful way of analysing this data, i.e. going beyond the sentence level, and looking at the relation between the idiom and its total embedding in the text. As an example, consider (27), where the text splits up what is clearly a coherent statement by Mr. Gort:

- (27) 'I'm ready to *bite the bullet*,' said Mr. Gort, one of two Commissioners who have said they would vote to increase fees. 'We're elected to make decisions.' (*NYT* 1996)

Many of the idioms examined in this study are simply dropped into the text, like stone into a well (such as *a dime a dozen*), but that is clearly not the entire story. As Naciscione notes, "Instantial use may be realized in one sentence or it may exceed sentence boundaries. It may also be carried over to the subsequent paragraphs, chapters, parts of a book or even cover a whole book". (Naciscione 2001: 30)

6. *WebCorp* and the Web

An attempt was also made to test various of these idioms on the Internet, using the *WebCorp* search engine (although this data was not included in the main body of the paper). Since there is no way of knowing how many Net hits one might maximally get, the data is primarily useful for confirming trends seen in the regular corpora. Here are some examples of collocations found via *WebCorp*:

- **Ace in the** {Albany, battlefield, collective, digital, economic, empty, energy, hair, headline, IT, Labour, money, plot, snook, video} **hole**.

- **Babe in the** {acting, advertising, Astral, baseball, cruciverbalist, Diplomacy, financing, historiography, Linux, NFL, ping, political, senatorial} **wood(s)**.
- **Baptism of/by** {budgetary, bureaucratic, cannon, chemical, fiscal, judicial, legal, lyrical, media, political, proxy, tactical} **fire**.
- **Drop in the** {budget, casino, celestial, conceptual, financial, fiscal, funding, income, linguistic, Microsoft, ratings, sexual, virtual} **bucket**.

Clearly, although a number are rather startling or obscure, the majority are relatively predictable, as well as tending to be one-word anchorings (as would be predicted by the right-branching nature of NPs in English). None recurred with any noticeable frequency, however. Perhaps the most startling result was that *WebCorp* rarely found the maximum number of hits it is set to report on, although that may have been caused in part by search problems with wild cards (we are after all not looking for the canonical form alone).

7. Conclusions

It seems clear that it is possible to split idioms and insert anchoring material, but that this is done rather more sparingly than was proposed (in pre-corpus days) by Ernst, who remarked “This sort of stylistic device is fairly common in newspaper and magazine articles” (Ernst 1980), and by Nicolas, who claims that “contrary to received views, at least 90% of V-NP idioms, including many usually regarded as completely frozen, appear to allow some form of (syntactically) internal modification” (Nicolas 1995: 233).⁹

As the present corpus-based figures indicate, ‘allowing’ a construction is not at all the same as ‘encouraging’ or ‘mandating’ it. At less than 3% overall, the anchorings remain a minor option, rather than a major pattern, and one that would also seem to be rather less frequent than Moon suggests, with her wealth of genuine examples. The frequencies found here clearly do not suggest that most of these idioms gleefully and repeatedly lend themselves to the verbal fireworks quoted in her Chapter 6; perhaps the *BNC* and such are merely duller. (But see the examples in the Appendix.)

Although the present study is basically synchronic (the data ranging from the early *BNC* citations to the 2002 *Independent*), a more diachronic study might well reveal shifts in the distribution of least some of these idiomatic expressions, particularly if they experience a surge in popularity such as the initial uses of *couch potato* (starting in the 1970s). Here, however, we lack a fundamental prerequisite, i.e. large amounts of easily accessible older corpus data. One might only note in passing that the data provided by the *Collins COBUILD Dictionary of Idioms* indicate that of the 55 idioms examined here, only two are labelled as “old-fashioned [BRITISH]”, and five more as “mainly British” or “mainly American”, while 29 are listed as relatively frequent (starred “key idioms”).

However, before dismissing this pattern of embedding as merely predictable and dull, let us examine a much more intriguing example from the

American poet and writer Richard Brautigan's *Trout fishing in America* (1967). Throughout the chapter called 'Trout Fishing in America Peace', he uses the title as an anchoring phrase of his own in a startling manner. First consider extracts from the opening of this chapter

In San Francisco around Easter time last year, **they** had a trout fishing in America peace parade. **They** had thousands of red stickers printed and **they** pasted them on their small foreign cars, and on means of national communication like telephone poles.
 The stickers had WITNESS FOR TROUT FISHING IN AMERICA PEACE printed on them. [...]
 They carried with them Communist trout fishing in America peace propaganda posters. (Brautigan 1967: 98):

Such a text has – to put it mildly – a rather strange pattern, one that is not found anywhere else in the work. To begin with, we may note that *trout fishing in America* is in itself a literary anchoring device that links this chapter's discourse with the many other uses of the phrase in the work. Next, remove *trout fishing in America* from all NPs where it occurs, such as *WITNESS FOR TROUT FISHING IN AMERICA PEACE*. If you do so, a text that could have been written by the John Birch Society steps forth, attacking the peace movement of the 1960s:

In San Francisco around Easter time last year, **they** had a [...] peace parade. **They** had thousands of red stickers printed and **they** pasted them on their small foreign cars, and on means of national communication like telephone poles.
 The stickers had WITNESS FOR [...] PEACE printed on them. [...]
 They carried with them Communist [...] peace propaganda posters.

If you then remove all the terms that are negatively evaluative of these people's activities, you find a simple report of what various people were doing in California at that time.

In San Francisco around Easter time last year, people had a [...] peace parade. They had thousands of [...] stickers printed and they pasted them on their [...] cars, and on [...] like telephone poles.
 The stickers had WITNESS FOR [...] PEACE printed on them. [...]
 They carried with them [...] peace posters.

TROUT FISHING IN AMERICA can thus be read as being inserted into the *WITNESS FOR PEACE* phrase to teach us to unhook ourselves from the propaganda hooks that ensnare the text.

In conclusion, you may well ask, *Where's the beef?* As far as our idioms go, it seems that they are best seen as not being the main attraction of the text they are embedded in, any more than a hamburger patty is the be-all and end-all

of every sandwich. And a good thing, too, as we otherwise really would have a beef!

Notes

- 1 Swedish runestone U 112. The runes are continuous, with single spacers between words; the larger spaces have been inserted here to make it easier to follow via the translation.
- 2 A fuller version: “The movie rights [to *A Confederacy of Dunces*] have been sold more than Uncle Tom, but now they’re in the hands of Drew Barrymore. God have mercy on her benighted soul – she’s going to appear herself as an expanded version of the stripper, Darlene, but today she committed further blasphemy and announced in a French interview that Ignatius will be played by Philip Seymour Hoffman.” (<http://www.worldwiderant.com/archives/001138.html>, accessed on April 10, 2003)
- 3 Nicolas remarks (1995: 248): “[...] although the research reported here was carried out before I read Ernst’s paper, it can reasonably be seen as an extension of his work [...].” Ironically, the present author had read Ernst’s paper, but not Nicolas’, before doing the research reported here, and the same caveat can thus be said to apply to the present paper.
- 4 For the purposes of this paper, it is assumed that newspaper CDs can function as corpora, a point that has been repeatedly argued elsewhere.
- 5 Although earlier newspaper CDs often are in machine-readable form (usually ASCII), more recent years tend to be in compressed files, so that only the search engine provided with the CD can be used to access the data. Additionally, titles and summaries are often in separate files, and only accessible when the relevant articles are displayed by the search engine.
- 6 To conserve space, only percentages are given, but the actual numbers are of course recoverable from the percentages and the number of tokens.
- 7 The figures with names included are virtually identical, except for *rain on your parade* (with names included, only 17.6%) and *birds of a feather* (with names included, only 7.5%).
- 8 In addition, a few, such as *out with the ark* and *blot on the escutcheon* (with only 5 tokens each) are so infrequent that they must be treated with caution.

- 9 At the end of his study, Nicolas concludes that “Almost all V-NP idioms can have syntactic NP modification. This is very clear, so that it is now apparently unmodifiable idiom NPs such as *make way* and *go halves* that stand in need of explanation” (1995: 249).

References

- Brautigan, R. (1967), *Trout fishing in America*. New York: Delta.
- Chaucer, G. (1977 [c. 1400]), ‘The Canterbury tales’, in: J. H. Fisher (ed.) *The complete poetry and prose of Geoffrey Chaucer*. New York: Holt, Rinehart and Winston.
- Collins COBUILD Dictionary of Idioms*, 2nd ed. (2002). Glasgow: HarperCollins.
- Ernst, T. (1980), ‘Grist for the linguistic mill: idioms and “extra” adjectives’, *Journal of linguistic research*, 1 (3): 51-68.
- Moon, R. (1998), *Fixed expressions and idioms in English: a corpus-based approach*. Oxford: Clarendon.
- Naciscione, A. (2001), *Phraseological units in discourse: towards applied stylistics*. Riga: Latvian Academy of Culture.
- Nicolas, T. (1995), ‘Semantics of idiom modification’, in: M. Everaert, E.-J. van der Linden, A. Schenk and R. Schreuder (eds.) *Idioms: structural and psychological perspectives*. Hillsdale (NJ): Lawrence Erlbaum.

Corpora

- British National Corpus*, world edition (2000). Oxford: Humanities Computing Unit.
- Broadcast News* CD-ROM (1995). Woodbridge (CT): Research Publications International.
- The Independent* CD-ROM (2002). Cambridge: Chadwyck-Healey.
- Los Angeles Times* CD-ROM (1990). Cary (NC): Dialog OnDisc.
- The New York Times* CD-ROM (1996). Ann Arbor: UMI.

Appendix: examples for further analysis

Since this paper has turned up a large number of interesting examples, it seems appropriate to provide some further data, both of obvious cases of anchoring, and of some borderline ones, all linked by idioms:

- (28) An energy economist, Charles M. Studness of Manhasset, said the secrecy surrounding the plans was troubling. ‘Are people trying to do something in the public interest, or is there **a hidden political agenda?**’ Mr. Studness asked. (*NYT* 1996)
- (29) The head of the Irvine Values Coalition, the group that pushed Measure N, said Plummer’s comments were proof that opponents would push what he called “the **hidden homosexual agenda**”. (*LAT* 1990)
- (30) In a city where running the schools has always been **a delicate political balancing act**, questions of bureaucratic authority also always concern political influence. (*NYT* 1996)
- (31) Dance: A **high-wire balancing act**; ENGLISH NATIONAL BALLET Corn Exchange London New Victoria Theatre Woking (where the dancers perform circus acrobatics). (*Ind* 2002)
- (32) Lingering is an Italian art, and it can be studied at Cova. Those at the bar manage **a particularly Italian balancing act** in which, standing, they are able to eat cake, drink espresso, gesture while speaking and keep their shawls and purses perched on their slender shoulders. (*NYT* 1996)
- (33) The pitcher says to the umpire: “You’re **blind as a baseball bat**, you eggplant, and your mother wears go-go boots”. (*LAT* 1990)
- (34) **Bones of Contention**: The Archaeopteryx scandals. (*Ind* 2002)
- (35) But Mr. Lebed clearly relishes his reputation as **a bull in the Kremlin china shop**. [...] ‘They think that they have lassoed me and that I have to obey and play by the rules,’ he said. (*NYT* 1996)
- (36) But with growing scientific sophistication, it is possible to learn much more about exactly what happens when carcinogenic chemicals rampage like **molecular bulls** through the body’s **cellular china shop**. (*NYT* 1996)
- (37) “rewriting” is probably too strong a phrase – the plan was transformed last week into the favoured infant of the Lord Chancellor who has, in his historic ceremonial capacity, been made to **carry the Great Can of State**. (*Ind* 2002)
- (38) DANA KENNEDY: I just blanked out for a minute. And Robert DeNiro who is a crack armed robber. And this three-hour story is basically **a cat-and-mouse game**, a **mano-a-mano** game, between the two of these guys in which Al Pacino basically tries to track down Robert DeNiro. (*BN* 1995)
- (39) KEITH BYARS has **one big chip on his shoulder pads** concerning JIMMY JOHNSON. (*NYT* 1996)
- (40) The flip side of Robin Leachifying the rich is that, given the basic human need to despise somebody, we focus on the poor. In a country with a **chocolate chip on its shoulder**, that sentiment has often gotten translated

- into racial terms (“Negroes are happier not reading,” old-time Southerners used to proclaim). (*LAT* 1990)
- (41) Such grungeplay is probably the strongest link to Velocity Girl’s new deal with Sub Pop: other than that, they hardly *fit the Bud-swilling, neck-thrusting bill*, and the world is probably a better place for it. (*BNC*: CK6 2483)
- (42) Instead, the man who passed his first coaching licence at the age of 22 when he was a young player at Celtic, has taken what many would consider to be *a managerial poisoned chalice*. (*Ind* 2002)
- (43) Ms. Schiffrin, like Mr. Dole and others who contend that Uncle Sam is strangling family life, *puts the tax-collecting cart before the income-generating horse*. Soccer moms and dads feel economic pressure these days not because the Government is taking too much out of their paychecks, but because employers aren’t putting enough in. (*NYT* 1996)
- (44) Title: ***PUTTING A PICKWICK CART BEFORE THE HORSE*** [...] Thirty years ago in medical school we studied the Pickwickian syndrome, named for the portly Dickens character, whose sufferers had poor sleep habits and daytime drowsiness. (*NYT* 1996)
- (45) Title: ON ***SKINNING SCHRODINGER’S CAT***
THOSE who follow physics the way other people follow baseball quietly cheered at the news last week that scientists had finally *skinned Schrodinger’s cat*. (*NYT* 1996)
- (46) Since the announcement of the layoffs, government officials and labor leaders have sought to *soften the economic blow* of the company’s departure. (*LAT* 1990)
- (47) I’m happy to stick to the original. And I’m determined to feed it to the undiscerning teenaged louts who hang around the family manse. Last time I tried, I decided to *soften the tannic blow* by mixing it with an equal part of orange juice. Reaction: mixed. (*Ind* 2002)
- (48) Wolves’ vociferous streak has a negative side, both Butler and Rae *tempting providence and the referee* after being cautioned. (*Ind* 2002)
- (49) picture and sound quality. Then Murdoch weighed in. ‘Why not sub-divide each channel into four or five channels,’ he asked. It’s called multi-plexing and it *upset the digital apple cart*. (*BN* 1995)
- (50) MILES O’BRIEN: Telescope dealer Martin Cohen says binoculars are a good first choice because if you’re not impressed by the stars, binoculars can easily be used for other things. But, if they *whet your astronomical appetite*, this might be the next logical step. (*BN* 1995)

NP-internal functions and extended uses of the ‘type’ nouns *kind*, *sort*, and *type*: towards a comprehensive, corpus-based description

Liesbeth De Smedt, Lieselotte Brems and Kristin Davids

University of Leuven

Abstract

*In this paper we investigate the various constructions containing one of the three main type nouns *sort*, *kind* and *type*. Basing ourselves on data from the COBUILD corpus and COLT corpus, we first present a subclassification of the main type noun constructions, which owes a lot to but also expands on Denison (2002) and Aijmer (2002). In comparison with the categories proposed in the current literature, we advocate finer distinctions mainly within the NP-internal uses of type nouns, by positing fundamental structural and semantic distinctions between head uses on the one hand and modifier uses (attributive and semi-suffix) and postdeterminer uses on the other. The subjectified qualifying uses and discourse marker uses of type nouns, by contrast, have been covered rather extensively in the literature. From the existing descriptions we retain the distinction between nominal, adverbial and sentential qualifiers, discourse markers and quotative markers. We then apply this descriptive framework to two British English data sets from opposing registers: written texts from the quality newspaper The Times (COBUILD subcorpus) and spontaneously spoken conversation between teenagers (COLT). The quantification of these analyses reveals strong asymmetries in the relative frequencies of the various type noun uses in the two data sets. While type nouns are used predominantly NP-internally in The Times, adverbial qualifiers and discourse markers predominate in the COLT-data.¹*

1. Introduction

In Present-day English, three nouns expressing the general meaning of ‘type’ are very frequently used: *kind*, *sort*, and *type*. These ‘type’ nouns may fulfil different functions in the NP, as in (1) and (2).

- (1) Funny Bones was based on the premise that there are two **sorts** of comedian. (CB – *The Times*)²
- (2) The problem was that the Bush administration said the funding should come from the cities and states. The cities and states said they didn’t have this **kind** of money. (CB – UK spoken)

The type nouns also have uses in which their meaning has been extended and modified via subjectification (Aijmer 2002, Denison 2002), as in (3) and (4).

- (3) She was lookin **kinda** dumb with her finger and her thumb in the shape of an “L” on her forehead. (www.charcards.com/AllStar.htm)
- (4) [...] those advert features that they do in the States, they’re like **sort of** like ... half hour adverts, like they’re like shows but they’re [...] (*COLT*)

The versatile behaviour of type nouns already attracted the attention of the great descriptive grammarians of the first half of the 20th century such as Kruisinga. In recent years, their relevance to the study of grammaticalisation and discourse patterns has made them into something of a hot topic in English studies.

However, despite the number of articles and papers devoted to type nouns, there is no agreement in the literature about a systematic, formally motivated classification of their different uses. The aim of this paper is to work towards such a comprehensive grammatical description. Assuming a functional approach (as in e.g. McGregor 1997), we will attempt to characterise the different grammatical functions in which type nouns occur in our data.

The structure of the article will be as follows. In Section 2, we will consider three grammatical analyses of type nouns which provide essential insights: Kruisinga (1932), Denison (2002) and Aijmer (2002). As we will see, it is in fact the NP-internal functions of type nouns that have been covered least well and that pose the greatest descriptive problems. By contrast, the uses of type nouns involving clear meaning shifts, for instance to ‘qualifying’ uses (Denison 2002), have already been covered better. In Section 3, we will set out our own descriptive framework, accounting for the different functional configurations found in our data. Finally, in Section 4, we will report on De Smedt’s (2005) quantified study of the NP-internal functions of type nouns proposed in Section 3, as well as of the extended uses of type nouns.

2. Grammatical classifications of type noun patterns in the literature

In this section we will discuss Denison (2002), Aijmer (2002) and Kruisinga (1932): three different basic grammatical classifications of type noun patterns. These classifications complement each other in interesting ways, because they invoke, partly, different types of formal evidence and because each author points out some patterns not considered by the others.

2.1 Denison (2002)

Denison (2002) posits two, or possibly three, basic constructions with *sort/kind/type of*, and some “semi-conventionalised variants” (2002: 3) of these basic patterns.

The basic constructions, which are distinguished on the basis of clusters of syntactic, semantic, formal and discourse features (Denison 2002: 2), are represented in Table 1. (N1 refers to the type noun and N2 to the second noun following *of*.)

Table 1: Denison's (2002: 3) classification of *sort/kind/type of*-constructions.

	<i>semantic head</i>	<i>discourse function</i>	<i>N1</i>	<i>N1 number</i>	<i>E.G.</i>
binominal	N1	discourse topic or anaphor	<i>sort kind type</i>	sg. or pl.	the sort of material
qualifying	N2	Hedge	<i>sort kind</i>	sg.	a sort of holiday
postdeterminer / complex determiner	?	anaphoric	<i>sort kind type</i>	sg.	these sort of skills

The first construction is called 'binominal' because the type noun functions as a full noun, which is the head of the noun phrase, while the *of*-phrase functions as a postmodifier to that head, e.g.

- (5) Collagen is the **sort** of material that is found already [...] in the dermis of the skin. (Denison 2002: 2)

According to Denison, N1 and N2 typically agree in number. In the 'qualifying' construction, *sort/kind + of* form a unit which hedges the categorial meaning of N2 for ironic or other purposes, as in

- (6) But I suppose it's as a that's as a **sort of** holiday, kind of doing you know nothing but sitting around (*ICE-GB*, quoted in Denison 2002: 2)

Only *kind* and *sort* (in their singular form) are used in this construction, and N2 is the head of the noun phrase.

The 'postdeterminer/complex determiner' construction has an uncertain status in Denison's current analysis. It is distinguished mainly on the basis of number incongruence between singular type noun and plural anaphoric determiner + plural N2, as in

- (7) I mean I don't associate you with uh you know one of **these sort of** skills like like driving. (*ICE-GB*, quoted in Denison 2002: 3)

Denison leaves it open whether the type noun or N2 constitutes the head of the noun phrase. More fundamentally, he raises the question whether this is really a distinct third pattern, or merely a re-analysis of the binominal construction with singular N1 + *of* + plural N2 (Denison 2002: 11).

The additional patterns identified by Denison are viewed intrinsically as variants of the basic constructions. Two of these are particularly productive. Firstly, there are several variants of the qualifying structure, in which there is no N2 and in which the *sort/kind of* string modifies adjectives or verbs, as in

- (8) I **sort of** saw his point. (*Frown*, quoted in Denison 2002: 3)

Denison calls this the ‘adverbial’ construction. The adverbial use of *sort of* has also given rise to a ‘semantically bleached’ discourse marker use, as in

- (9) As I remember it used to be **sort of** like fairly common for a Tuesday [...] (*ICE-GB*, quoted in Denison 2002: 4)

Secondly, there is what Denison (2002: 4) refers to as the ‘semi-suffix’ use of *type (of)*, as illustrated in (10):

- (10) what you’re saying is we need multiple **type of** I mean ideally we need a multiple type building [...] sorry a building with multiple type rooms. (*LLC*, quoted in Denison 2002: 4)

We can conclude that Denison distinguishes two basic sets of type noun-constructions. On the one hand, there is the ‘binominal’ construction in which the type noun is used in its original sense of ‘subclass’. On the other hand, there is the ‘qualifying’ construction in which *sort/kind of* has a hedging function; this pattern has led to the adverbial construction and the bleached discourse use.

2.2 Aijmer (2002)

Using different labels, Aijmer (2002) draws a distinction between different type noun uses, which is similar to the fundamental distinction made by Denison between the binominal use and the qualifying use.

On the one hand, she distinguishes the basic NP structure consisting of determiner + type noun used as head + postmodifier, [*a [sort/kind/type [of NP]]*], as in *A robin is a sort of bird*, which, she notes, encodes the meaning ‘X is a hyponym of Y’.

On the other hand, the strings *sort of* and *kind of* fulfil a great variety of what are called ‘particle’ uses by Aijmer (2002: 176). She (2002: 180) views these particle uses as the result of grammaticalisation, pointing to formal evidence such as the bonding of *of* to the type noun and its phonological reduction as reflected in spellings like *sorta/kinda*. Their different meanings come about through processes of pragmatic enrichment (Traugott 1995, 1996), in which the literal ‘type’ meaning has shifted to a hedging meaning and further, through subjectification, to an affective discourse particle. In the particle uses, the scope of *sort of/kind of* has also extended to other syntactic environments than the NP, such as adjective phrases, verb phrases, and even whole sentences.

Within the particle use, Aijmer makes a general distinction between the evidential and the affective function. As an ‘evidential’ marker, *sort of* functions as “an adjuster word, as an indicator that the following word or construction functions on a different level of talk (the meta-level *sort of*), as an indicator of number approximation, signalling a lexical gap or lexical imprecision or as a self-repair signal” (Aijmer 2002: 192). The meta-level *sort of* subsumes a quoting use,

introducing reported speech or onomatopoeic phrases (11), as well as a use marking a "special idiom" in front of "technical, rare, foreign, formal, vulgar, idiomatic words" (Aijmer 2002: 192), or ad hoc invented phrases. In all cases the meta-level *sort of* signals that the following word or phrase has a special status or does not belong to the regular vocabulary of the speaker, as in (12).

- (11) [...] was still – pretty wealthy – I've never seen a sort of bottle after bottle, **sort of** pop pop popping all the time and everybody got awfully drunk I remember (*COLT*, quoted in Aijmer 2002: 186)
- (12) Sorry I didn't mean to I didn't mean to sound **sort of** prissy (*COLT*, quoted in Aijmer 2002: 195)

The second major particle use is the 'affective', or interpersonal, use, in which *sort of* mainly serves to hedge the illocutionary force of utterances in order to avoid disagreement with the hearer and claim common ground (Aijmer 2002: 191). Further submeanings include densifying/downtoning, compromising, signalling intimacy between the speakers or hedging of strong opinions as strategies of positive and negative politeness respectively (Aijmer 2002: 202–207).

In conclusion, the main contribution of Aijmer's study clearly lies in her fine-grained description of the various particle uses of *sort of*, observed in her data, which she accounts for within a coherent framework of grammaticalisation, semantic shift and pragmatic enrichment.

2.3 Kruisinga (1932)

Contemporary descriptions of type nouns seem to assume head status for the semantically more neutral type noun uses in the NP, as in Denison's (2002) 'binominal' construction. It is only for uses which have undergone a clear semantic shift, such as Denison's (2002) 'qualifying' construction, that a concomitant grammatical shift from head to modifier is posited.

By contrast, Kruisinga (1932), in his earlier treatment of type noun uses, did not only discuss their nominal head use as well as their ability to function "as adjuncts to verbal forms [...], and to adjectives" (Kruisinga 1932: 399), but he also identified non-qualifying modifier uses in nominal groups. Whereas in head uses the type noun functions as an "independent element of the group" (Kruisinga 1932: 395), it is "entirely subordinated in meaning" (Kruisinga 1932: 395) in modifier uses such as

- (13) What **sort of** a man is he to see? (*The strange case of Dr Jekyll and Mr Hyde*, by Robert L. Stevenson 1886, quoted in Kruisinga 1932: 178)

- (14) What **sort of** weather are we going to have? – It doesn't look very promising at present, but you never know! (*Spoken English*, by Collinson, quoted in Kruisinga 1932: 178)

In these examples, Kruisinga (1932: 178) notes, *what sort of* is used to inquire about the quality of persons or things. In other words, *what sort of* asks the hearer to attribute a quality to the entity designated by the noun following *sort of*, such as 'weather' – 'not very promising' in (14).

In correlation with his description of semantically subordinated uses of type nouns, Kruisinga also discusses a number of grammatical corollaries of their modifier status. The shift from head to modifier status of type nouns is made possible by the special nature of the preposition *of*, which "can sometimes make a preceding noun (instead of the following noun) into an adjunct" (Kruisinga 1932: 391). Kruisinga points out these two possible structural analyses for nominal expressions with both 'species' nouns and 'measure' nouns (such as *dozens of*).³

In general, Kruisinga (1932: 397) points out, "the subordination of *sort*, *kind*, *manner*" also influences their stress. As a modifier the type noun is typically only medium-stressed (Kruisinga 1932: 397). Another formal characteristic of the modifier status of type nouns is that adjectives in front of the type nouns have scope over N2 (i.e. the noun in the *of*-phrase) (Kruisinga 1932: 397), as in

- (15) But Kezia bit a big piece out of her bread and dripping, and then stood the piece up on her plate. With the bit out it made a dear little **sort of** a gate. (*Bliss*, by Katherine Mansfield 1920, quoted in Kruisinga 1932: 397)

Number incongruence between the type noun and the determiner preceding it is also taken as a general sign of modifier status of the type noun by Kruisinga (1932: 398), as in (16).

- (16) It is a charming talent: all **manner of** arts and graces proceed from it. (*The Times*, 29 August 1913, quoted in Kruisinga 1932: 398)

Finally, Kruisinga (1932: 396) regards the presence of an "indefinite article before the prepositional noun" as an indication of head noun status of the second noun, as in (17).

- (17) He is a good **sort of** a fellow after all. (*The last chronicle of Barset*, by Anthony Trollope 1867, quoted in Kruisinga 1932: 396)

In conclusion to this survey of the literature, we can say that there is general agreement on the fact that the qualifying use of type nouns involves a shift to modifier status. However, only Kruisinga unhesitatingly posits a head to modifier shift for some non-qualifying uses within the NP, such as the 'attributive' use of type nouns.

3. Classification of type noun functions

In this section, we will propose a synchronic characterisation of type noun functions that covers the main semantic and formal fault lines in our data. The general semantic value of these functions will sometimes be related to more specific discourse patterns, and the formal features of these functional categories include grammatical structure and prosody, as well as lexical selection restrictions.

The corpus data include, firstly, general data sets of exhaustive extractions of *kind(s)/sort(s)/type(s)/of/-a* from *The Times* subcorpus of the *COBUILD* corpus and from the *COLT* corpus, compiled by De Smedt (2005). These sets can be assumed to be representative of the various type noun uses in formal written British English on the one hand, and in informal spoken British English on the other. The relative frequencies of type noun functions found in them will be discussed in Section 4. For descriptive and illustrative purposes, reference will also be made to a further data set compiled by Lavrysen et al. (2005), who extracted from the whole *COBUILD* corpus all uses of *kind(s)/sort(s)/type(s)/of/-a* pre-modified by adjectives. As we will see, these data are particularly enlightening with regard to non-qualifying modifier uses.

3.1 Head use

As seen in Section 2, the fact that type nouns can function as head is not in question. What this section sets out to do is to define the head use more narrowly, and delineate it *vis-à-vis* the (non-qualifying) modifier and postdeterminer use. To this end, the head use's structural properties will have to be identified and its semantics and most general discourse uses will have to be specified.

If the type noun functions as lexical head, the whole NP is concerned with designating (sub)classes. Mostly, in these NPs, *of* + N2 specify the more general class that the subclasses referred to by the NP are included in, as in (18):

- (18) There are 5 **types** of animals you can have on your farm: your dog, horse, cows, chickens, and sheep. (www.fogu.com/hm4/moo/index.htm)

but there may also be no N2, as in (19):

- (19) All **types** (except _{file}) can be converted to and from _{string}. (www.brics.dk/bigwig/refman/types/#conversion)

Crucially, if qualitative or classifying adjectives occur before the type noun, they apply to the subclass:

- (20) In 1981 a doctor found 5 previously healthy young men with *Pneumocystis carinii* pneumonia, **a very rare type** of infection. (*CB* – UK ephemera)
- (21) George Broadhead suggests a gay alternative to weddings and reckons there is little likelihood of **the Danish sort** of reform being introduced in Britain in the future (*CB* – UK magazines)

The intonation going with head use has, as noted by Aijmer (2002: 176-177), either primary stress on the type noun (or its determiner or premodifier) or on N2. It seemed to us that in our data⁴ these two basic options for primary stress correlated with a contextual emphasis either on the subtype or on the superordinate. Thus, corresponding to the two possibilities of primary stress, two types of discourse context could be distinguished.

Firstly, NPs with type noun in head position can be used in contexts in which (aspects of) the subclass are presented as new. The speaker may identify the relevant subclass as such, as in (21) above. In this example, some superordinate category such as ‘alternatives to conservative policies’ is given in the discourse, and the NP with type noun brings a new subtype of that superordinate type into the discourse, viz. *the Danish sort of reform*. It may also be quantitative or qualitative aspects of the subclass that are presented as new, and hence marked by primary stress. Thus, the speaker may focus on communicating how many relevant subclasses there are, as in

- (22) *Funny Bones* was based on the premise that there are **two sorts** of comedian. (*CB* – *The Times*)

Alternatively, the speaker may focus on an attribute of the subclass, as in (20), which stresses that *Pneumocystis carinii pneumonia* is *a very rare type of infection*.

Secondly, NPs with type noun as head can be used in contexts in which the superordinate category is new, as in Denison’s example (5) above and (23).

- (23) The late-17th century console has an elaborate top inlaid with agate and *pietra paesina*, **a rare type** of marble which in its colours and veining looks like a *paesaggio* or landscape. (*CB* – UK magazines)

These examples instantiate the hyponymic use pointed out by Aijmer (2002: 176). For instance in (23), the (rather obscure) class of *pietra paesina* is categorised by *a rare type of marble* as belonging to the (better known) general class of marble.

Whether the speaker focuses on the subtype or on the superordinate type is a discourse effect – it does not change the structural relation between the type noun in head position and postmodifier *of* + second noun. It is in virtue of this modification relationship that the structure type N + *of* + N2 designates specific

subtypes of more general types. As such, these structures clearly fall within what Bache (2000: 239) has referred to as the 'categorisation' functional zone in the NP, from which, preminally, the 'determination' and 'modification' zone have to be distinguished. As pointed out by Langacker (1991: Ch. 2), the categorisation elements in the NP contribute to the 'type specification': they designate a mere type. This type meaning is obvious in the rare grammatical contexts where English uses the 'categorisation' unit of the NP only, as in compounds such as *dog lover*, which designates a lover of the type, not of one specific, 'dog' (Langacker 1991: 75). It is the full NP which designates a specific instance, identified in function of its position *vis-à-vis* speaker and hearer. For instance, *my very own dog* refers to a specific pet, the one, as indicated by the determination unit *my very own*, owned by the speaker. To the specific instance referred to by the combination of 'categorisers' and 'determiners' in the NP, qualities and specific features may be attributed by 'modifiers', as in *my beautiful golden dog*.

In NPs with type nouns as head, the type noun + *of* + N₂ designates the categorisation, with which the determiners interact to single out a specific referent. Because the head noun designates a subclass, such NPs have either generic reference, as in *a rare type of marble* in (23), or generalised reference, as in (24).

- (24) Labour has always accommodated **different kinds** of socialist and should continue to do so. (*CB – The Times*)

The former refers to a true species or subclass of marble, whereas *different kinds of socialist* in (24) makes generalisations over concrete members of the Labour party (see Willemse 2005: 188-192, for clarification of the difference between generic and generalised reference). Some of these head uses, while overtly coding generic or generalised reference, may *imply* reference to specific instances of the classes in question, as in (25), which discusses instantiations of lust, guilt, etc. in the stage play 'Herod'.

- (25) Berkoff is Herod and Zigi Ellison is the eponymous temptress. There will be lust, guilt, incest, lewdness, licentiousness and **a dozen other kinds** of sick amorality. (*CB – The Times*)

Ward and Birner (1995: 732) have characterised the mechanism by which reference to a concrete instance is implied by a NP whose overt reference is general or generic as 'dual reference'.

3.2 Modifier uses

In agreement with Kruisinga (1932), we hold that there are type noun constructions in which the type noun has clearly been demoted from head to modifier status, as indicated by the semantics and formal characteristics of these constructions. Within the modifier uses of type nouns a further distinction will be made between attributive and semi-suffix modifier uses.

3.2.1 Attributive modifier use

When the type noun functions as modifier rather than head of the NP, it is always preceded by an item invoking a quality or specific feature. This item may be lexical, and is then usually an adjective (26), or it may be a function word such as an interrogative determiner (27).

(26) Being **an accommodating sort of** bloke, he let me take the car around the paddock at Silverstone (*CB* – UK magazines)

(27) **what sorta** question is that?! I assume You're a beginner...
(www.menshealth.co.uk/talk/thread.phtml/post740523/-63K)

Crucially, the qualities named or inquired about apply to an instance of N2, not to the type: the modifier uses of type nouns attribute these qualities to the instances in question. How and why they do differs in cases like (26) and (27). We will first look more closely at the cases with premodifying adjectives and then at those with interrogative determiners.

The type nouns in examples such as (26) are clearly cases of “phrases of the form *kind/sort of* N [...] tak[ing] premodifiers plainly related to N rather than *sort*, both in semantics and in concord” (Quirk et al. 1972: 930). This ‘transferred’ (Halliday 1985a: 174) use of the adjective involves a true semantic and structural re-analysis. In contrast with NPs with type noun as head, which have generalised or generic reference (as in *a very rare type of marble*), the adjective in (26) applies to a concrete instance of the category ‘bloke’. With reference to Bache’s (2000) functional zones, *a very rare type of marble* is parsed as [determination: *a*] [modification: *very rare*] [categorisation: *type of marble*], but *an accommodating sort of bloke* has to be analysed as [determination: *an*] [modification: *accommodating sort of*] [categorisation: *bloke*]. In this context it is interesting to reconsider the variant of the modifying use of type nouns discussed by Kruisinga (1932: 396) with indefinite article in front of N2, such as *what sort of a man* and *a good sort of a fellow* in examples (13) and (17) cited above. Kruisinga saw the article in front of N2 as an overt indication of the head status of N2. In our view, it also overtly signals that an instance of N2 is being referred to. Lavrysen et al. (2005) found that the variant with *a(n)* in front of N2 is now rather uncommon in English usage. However, in principle this variant is still possible with the modifier uses under discussion here, as shown by (28a) and (28b), but ruled out with NPs in which the type noun functions as head (29).

(28a) Only last year, Richards branded Australian team manager Bobby Simpson a bad loser and **a rather sour sort of** guy (*CB* – *Today*)

(28b) I’m **an ordinary sort of** a guy. (*CB* – UK spoken)

(29a) She was turning into **the worst kind of** Victorian husband. (CB – *The Times*)

(29b) *She was turning into **the worst kind of a** Victorian husband.

Thus, the variant with *a(n)* in front of N2 can be used as a test to recognise modifier uses of type nouns, provided N2 is a singular count noun.

Phonologically, the modifier status is reflected by the fact that *sort/kind/type + of* in these uses never have primary stress. It is the premodifying adjective that receives the primary stress. The next stressed element in the NP is N2. In between these two stressed elements, type noun + *of* are non-salient, with the type noun not stressed and *of* typically reduced to /ə/. In informal registers, this reduction of *of* may be reflected by spellings such as *typa* and *sorter*.⁵ Semantically, type nouns in modifier position no longer mean '(sub)class', but more something like 'qualitative variant'.

Lavrysen et al. (2005) investigated the lexical sets associated with adjective and N2 position in these constructions in an exhaustive extraction from the COBUILD corpus on adjective + type noun + *of* + N2. They found that specific selection restrictions are associated with both sets, and that this modifier use of type nouns also tends to be associated with specific motives in Present-day English. By far the most prominent motif is that involving an adjective attributing a character trait to a person, as in *an accommodating sort of bloke, clubbable kind of politician, a pragmatic kind of guy, inspiring sort of person, a very dramatic sort of person, a liberal, open-minded kind of daughter, a verbal type of poet, a scattered sort of person, a paternalistic sort of chap*. This motif seems to have a fairly long ancestry, as it already crops up in example (17) above from Trollope cited by Kruisinga (1932: 396). More generally, Lavrysen et al. (2005) observed that the adjectives premodifying type nouns in modifier position often are rather unusual and infrequent adjectives, e.g. *a niggardly sort of word, a pretty torrid sort of programme, a peaceful sort of sorrow, in a moody-broody sort of way*, etc. These lexical patterns suggest that this construction, with its rather limited application in Present-day English, is veering towards an interpersonal, meta-linguistic, use, in which the speaker may, for instance, be apologetic or ironic about an unusual choice of adjective. In the process, *sort/type/kind + of* increasingly acquire the feel of a fixed string appended to the adjective, and, in this way, seem to be gradually demoted from core modifier status to a more enclitic position.

We now turn to the second pattern of modifying uses of type nouns, in which the type noun is preceded by a determiner, mostly interrogative *what*, as in examples (30), (31), (32) and (27) above.

- (30) You go in there and you free people and you give them the right to choose **what sort of** a government regime they're going to have and they choose the wrong one. (www.signposts.org.au/index.php/archives/2003-04)
- (31) "Why doesn't your child live with you? Why couldn't you make him happy? **What sort of** mother are you anyway?" (www.questia.com/PM.qst?a=refresh&docId=0899319&type=book)
- (32) **What kinda** jerk would I be if I simply turned my back on her? (www.bigblack.blogspot.com/2005_01_01_bigback_archive.html)

In (30), *what sort of* is used in its descriptive meaning of 'what qualitative variant of'. As also signaled by *a* in front of N2, *what sort of* applies to an instance of the category 'government'. It functions as an instruction to attribute a quality to the instance of N2. However, in examples like (31), (32) and (27), *what sort/kind of* function differently. Rather than instructing the addressee to replace *what sort/kind of* by a specific descriptive attribute, they have undergone (different degrees of) subjectification. Bolinger (1972: 32) offers a sharp interpretation of these semantic shifts from 'describing' to 'evaluating' and 'intensifying'.

In the first stage, illustrated by (31), the speaker decries something by pretending that it does not deserve its name (Bolinger 1972: 32). In (31) it is implied that the addressee is a bad or unworthy instance of the category 'mother'. Via metaphorisation, this use develops further into an intensifying use (Bolinger 1972: 32), as in (32), where *kinda* further emphasises the negative features that are part of the semantics of *jerk*. In uses with neutral nouns such as *question* in (27), finally, *what sort/kind of* seems to have acquired a certain negative value of its own. Interestingly, the variant with *a* is still possible with the 'implicational' use in (31) – *what sort of a mother are you* –, but seems forced or even impossible with the intensifying use in (32): *what kind of a jerk*. In any case, in our classification, all of the uses of *what sort/kind of* exemplified by (30), (31), (32) and (27), which display various stages in the process of subjectification, are ranged with the 'modifier' uses of type nouns.

3.2.2 Semi-suffix use

This section is concerned with the semi-suffix use of type nouns, whose core cases are readily recognizable, and are illustrated by examples (33)–(39):

- (33) Listen, we have had sudden employment in the nature of developing a European-**typ**a film. (www.hexmaster.com/goonscripts/s08e16.pdf)
- (34) He wears a bright red superhero (**kinda**) costume with white-and-black sneakers. (www.whatacharacter.com/g-o/n--age1.htm/)

- (35) It is an iMac **sorta** phone. (www.ciao.co.uk/Reviews/Motorola_V60_5297096)
- (36) It's a Spielberg **Kinda** Christmas. (www.netribution.co.uk/features/carnal_cinema/96.html)
- (37) ..., it's a blink thrice or you'll miss it **kind of** town, ... (*CB – The Times*)
- (38) Martin has written at length in *The New Yorker* about Andrew Motion's biography of Philip Larkin: 'A policeman-in-the-head **kind of** book. (*CB – The Times*)
- (39) this is why im crazy about u. in a non-'worship and serve me forever' **typa** way ;-) (www.7thrimofhell.blogspot.com/1990/01/blog-post.html)

The semi-suffix use of type nouns has a number of *prima facie* similarities and dissimilarities with the attributive modifier use. As for similarities, in both, type noun + *of* is preceded by a linguistic expression, which receives primary stress. This expression applies to N2, not to the type noun.

There are also noticeable differences. In examples (33)–(39), the position in front of the type noun string is not associated with the class of qualitative adjectives. Rather, we find classifying adjectives such as *European* (33), nouns such as *superhero* (34), brand names such as *iMac* (35), proper names such as *Spielberg* (36), and longish nonce expressions, which may then be hyphenated or surrounded by quotation marks, as in *policeman-in-the-head* (38), and *non-'worship and serve me forever'* (39). Most of this material is not readily parsed as a premodifier subordinated to the type noun: rather, it is the type noun string which is felt to submodify this material, and thus to function as an enclitic. The practice of sometimes linking the type noun with a hyphen to the characterisation preceding it signals its clitic-like nature particularly clearly. Frequent as such expressions followed by *sorta/kinda/typa* are on the Internet, the impression may be created that informal registers use this spelling variant very consistently for the semi-suffix use. There is no doubt that this widespread practice reflects the language users' awareness that, phonologically, the type noun does not receive stress and that *of* is reduced in this use. By the same token, as already noted in endnote (4), this spelling practice is not restricted to semi-suffix use – it is even found with head use in informal registers – and it is not criterial to it, as is shown by examples (37) and (38).⁶ Hence, the *sorta/kinda/typa*-spelling should not suggest that the semi-suffix use of type nouns is easy to delineate – or wholly unrelated to the attributive modifier use of type nouns discussed in the previous section.

We tentatively propose the following functional definition of the semi-suffix use category: the lexical material preceding the type noun string is typically a classifying (rather than qualitative) element, and the type noun string itself always has some sort of interpersonal (rather than purely descriptive) value. If we

consider examples (33)–(39), we see that not only adjectives such as *European*, but also the other expressions in front of the type noun appear to be ‘classifiers’ (Halliday 1985a: 164), i.e. pronominal modifiers indicating a subtype of the general type designated by the head noun (N2). However, particularly the more creative subclassifications may imply qualities: for instance, *policeman-in-the-head* in (38) probably suggests qualities such as ‘inquisitive’ and ‘judgmental’. Moreover, the type noun strings always ‘frame’⁷ the categorisation preceding it in an interpersonal way. They have a general metalinguistic value and may further convey various interpersonal values such as:

- hedging the classifier, as in (34), where the label used may not be fully correct or applicable;
- downtoning of what might seem a high-flown label, such as perhaps in (33);
- humour and irony, as in the extravagant and farfetched nonce-expressions in (37) and (39).

In our view, the phonological and structural ‘demotion’ of the type noun string naturally ‘goes with’ the expression of such interpersonal values. The categorisation is foregrounded, but at the same time it is presented from a specific subjectified angle.

In the light of these observations about the semi-suffix use, it turns out to have more affinities with the attributive modifier use of type nouns than appears at first sight. Both involve unusual lexical items and the type noun strings following them have a metalinguistic flavour. Hence, even though both these features tend to be pushed further in the semi-suffix use, the boundaries between attributive modifier and classifying semi-suffix use are fuzzy, as illustrated by a set of examples such as *a quirky kinda song*, *a bluesy sorta song*, *a luv tyra song*. Consequently, it seems reasonable to consider attributive and semi-suffix uses as the two ends of one internally graded category. In our view, the hypothesis should also be considered that the semi-suffix use of type nouns derives diachronically from their attributive modifier use.

3.3 Postdeterminer use

In Halliday’s (1994) functional description of the English NP, the postdeterminer is defined as a ‘secondary deictic’ which adds ancillary information to the basic meanings ‘identifiable’ versus ‘non-identifiable’ referent, expressed by the primary determiner in the NP. It has been argued by Breban (2002 and forthcoming) and Breban and Davidse (2003) that it is probably better to think of determiners and postdeterminers as together forming ‘determiner complexes’, which can express more complex identifiability statuses and phoric relations than simple determiners. For instance, a determiner and postdeterminer complex such as *another* (whose functional unity is also reflected in the orthography) can express the introduction of a new instance of a discourse-given type, as in

- (40) Our people estimate that half the money was lost. We don't want **another** big problem of fraud and abuse. (CB – UK spoken)

In this way, as observed by Martin (1992: 116), a NP whose primary reference value is 'non-identifiable', may still incorporate an anaphoric relation, viz. to the instance of 'fraud' described in the preceding discourse. Regarding adjectives used as postdeterminers, it has been pointed out by Breban (2002 and forthcoming) that they do not have the meaning of attributing a quality to the instance referred to, but express, together with the determiner, the identifiability status of the referent in the discourse and its relation to other referents. For instance, whereas *another* in (40) indicates that a 'new' instance of fraud and abuse is being referred to, besides those implied in the previous discourse, predicative *other* in (41)

- (41) Everybody's trying to be **other** than what they really are... (CB – US books)

refers to the qualities and personality traits of *everybody* (see also Huddleston and Pullum 2002: 1145).

To quote another example of determiner and postdeterminer complex, *the same* in (42) expresses reference to a 'generalised' concept, abstracting one concept of shape and strength from the two instances talked about.

- (42) He agreed that at 85kg, his normal weight, he is not heavily-built for a rugby player. Physically, we are not **the same** shape or strength as British or Australian rugby players. (CB – Australian newspapers)

According to the *Oxford English Dictionary* (1933, vol. 9: 74-75), *same*, which derives from Indo-Germanic **some-*, meaning 'level', came into English as a postdeterminer adjective. Throughout the history of English, it has always formed a fixed unit with definite determiners.

It will be argued in this section that type nouns can also be used to form determiner complexes with determiners. As with adjectives used as postdeterminers, specific meanings of the type nouns ancillary to the expression of identifiability statuses and phoric relations are activated. We will begin by considering a number of examples in which the determiner complex expresses an **anaphoric** relation.

- (43) Comics from the South of England tended to be brash and outgoing, impressing the audience with how clever they were. In the North, **this sort of** style was not at all popular.
(http://pages.britishlibrary.net/mikepymm/george_formby_snr.htm)

- (44) ... a man I can trust and believe in. It takes time to find **that sort of** man, but he is worth the wait.
(www.forums.plentyoffish.com/16073117datingPostpage4.aspx/)
- (45) And to me Orlando Bloom is trying to be Johnny Depp, I mean they both have **the same sort of** looks.
(www.whimsical-strawberries.set/archives/00000049.html)
- (46) The purpose of ‘similar goods’ in the 1994 Act and EC Directive was to provide protection and separation for **a similar sort of** penumbra.
(*CB – The Times*)
- (47) She said: “I think the girls are amazing. There’s no way I could muster up **that kind of** energy during a Texas gig. I’d be knackered for a week.”
(*CB – The Sun*)
- (48) The problem was that the Bush administration said the funding should come from the cities and states. The cities and states said they didn’t have **this kind of** money.
(*CB – US National Public Radio broadcasts*)
- (49) A pet owner could bury a wire carrying a radio signal which activated the collar device when the animal approached the boundary, he said. Dr. Blackshaw said barking inhibitors were usually only effective in 25 percent of cases. Using **these sort of** devices means you are treating the symptom rather than the cause she said. (*CB – Australian newspapers*)
- (50) We were only able to respond that we were unaware of any evidence linking plastic milk bottles and cancer, but your investigative report puts the whole issue into perspective. Unfortunately **these sort of** scare tactics do a lot of harm. (*CB – Australian newspapers*)

In (43), *this sort of style* refers to a concept of ‘style’ which generalises from the instance discussed in the previous sentence. *This sort of* thus expresses both the reference to a generalisation of ‘good style’ and its anaphoric relation to an instance of it discussed in the previous sentence. In (44), *that sort of man* is likewise anaphorically related to the preceding description *a man I can trust and believe in*, and refers to an imagined person of that description who as yet has not materialised in the writer’s life, i.e. to what is traditionally called ‘a non-specific’ referent. The referential meanings conveyed by *that sort of* in this example can also be expressed by a NP with predeterminer *such*: *It takes time to find **such a** man*. (For further discussion of the affinities between demonstrative and type noun + *of* and *such*, see Denison 2002: 6 and Mackenzie 1997: 89).

In (45), *the same sort of looks* expresses reference to a generalisation over the looks of two individuals mentioned in the previous clause, Orlando Bloom

and Johnny Depp. This example is very comparable to (42), which has *the same shape and strength*, but in (45) *sort of* nuances the identity claim somewhat. In (46), *a similar sort of penumbra* introduces a new instance of the discourse-given type 'penumbra' (as does *another* in (40)), with *similar* stressing the similarity to previous instances. As in (45), *sort of* somewhat relativises this similarity. The presence of postdetermining adjectives *same* and *similar* in these examples contradicts Denison's (2002: 3) claim that the postdeterminer use of type nouns is incompatible with other postdeterminers.

In (47), *that kind of energy* construes an anaphoric relation to the instance of energy associated with the girls' performance in the preceding discourse, as well as reference to a more general concept of energy. However, whereas in the previous examples this generalisation invoked mainly the qualitative specifications of the concepts in question, *that sort of energy* activates the size implications of 'energy'. There also seems to be some intensification involved in this use of *that kind of*, which conveys 'a lot of'. The frequently used expression *that sort of money*, exemplified by (48), can be given a similar description; the intensification of the quantitative concept is even stronger here, as it conveys 'so / too much'.

Examples (49) and (50) illustrate the use of *sort of* and *kind of* as postdeterminers in plural NPs. The anaphoric relation to instances in the previous discourse, and the generalisation of reference to higher-order concepts named by *devices* and *scare tactics* can again be observed. As pointed out to us by Peter Willemse (personal communication), plural NPs with *these sort/kind of* also often imply a 'broadening' of reference. For instance, in *these sort of scare tactics*, the speaker includes not only the examples previously mentioned but also any other instances covered by that generalisation.

Besides the anaphoric uses illustrated in (43)-(50), type nouns are also often used in determiner complexes pointing forward in the NP, as in

- (51) If the war reignited, it could spread and spark **the kind of** conflict that has drawn Americans into two larger wars this century. (*CB – The Times*)
- (52) ... so it might as well go to Richard Holbroke. It was he who pitched together the Dayton accord, bringing to Bosnia **the kind of** peace Stalin and Hitler brought to Poland. (*CB – The Times*)
- (53) I position myself in the same way and I also try to use **the same type of** language as they do. This sends out a subconscious message that I understand how they feel. (*CB – The Times*)
- (54) We have spent a great deal of money trying to keep our players out of the reach of predators, but there was no way we could match **the kind of** offer made to Craig from our own resources. (*CB – The Times*)

In these examples, the NPs with postdetermining type noun refer cataphorically to concrete instances described by the restrictive relative clause in the postmodifier, and generalise from these. As with the anaphoric examples discussed above, some of these can be paraphrased with predeterminer *such*, e.g. (52) *such a peace as Stalin and Hitler brought to Poland*. In these cataphoric examples, the determiner complex with type noun can also interact either with qualitative specifications (e.g. (52)) or with size implications (54) of the concepts expressed by N2.

Having looked at the postdeterminer uses of type nouns illustrated by (43)–(54), we are now in a position to compare the category proposed by us with Denison's (2002) approach to the postdeterminer construction. As we saw in Section 1, Denison (2002) tentatively posits a postdeterminer construction of type nouns, the main formal evidence of which is the number incongruence between plural determiner and singular type noun, as illustrated in (49)–(50). He wonders, however, whether this construction is not just a re-analysis of the binominal construction.

We have approached the postdeterminer uses of type nouns functionally, noting the parallelisms with postdeterminer uses of adjectives. In a first step, this led to discussion of the referential values expressed by determiner complexes involving type nouns, as observed in the examples above. We will now further argue why we view the postdeterminer use of type nouns as a distinct construction, different from the head use of type nouns both in semantic and grammatical terms.

As noted in Section 3.1, NPs with type noun as head are concerned with discussing subclasses and hyponymic relations between classes. They all operate in what Langacker (1991) has referred to as the 'type universe' of interpretation, making statements such as 'There are *n* subclasses *T* of class *Y*', '*t* is an instance of subclass *T* which is a hyperonym of class *Y*', etc. Type nouns used as head followed by *of* + N2 all have generalised or generic reference in some way. This reference is realised by the interaction between determination and categorisation, as reflected in the parsing of 5 *types of animals* in (18), as [determination: 5] [categorisation: *types of animals*].

By contrast, type nouns used as postdeterminers construe, together with determiners, relations of anaphora or cataphora (cf. Denison 2002: 3) to concrete instances and generalise from them. These referential meanings are realised by the unit consisting of determiner and postdeterminer: an example such as *the same sort of looks* in (45) thus has to be parsed as [determination: *the same sort of*] [categorisation: *looks*].

The formal realisation of these referential meanings is not restricted to NPs with number incongruence. As shown in the semantic discussion of examples (43)–(54), the construal of anaphoric and cataphoric generalisation relations is not restricted to NPs with *these/those sort/kind of*. However, number incongruence does characterise plural NPs with demonstrative determiner and postdetermining type noun. This so-called 'number incongruence' reflects the fact that the number of such NPs is determined by N2, not by the type noun, as the

latter has coalesced with the determiner to express a referential and phoric meaning comparable to *such* (*a*). As argued by Breban (2002 and forthcoming), this coalescence is indicative of the fact that postdeterminers tend to derive from lexically full, more autonomous units such as attributive adjectives – or head nouns in the case of type nouns – via a grammaticalisation process (Hopper 1991, Lehmann 1985). In this process they shift towards general referential meanings and are formally bound to the determiner. NPs with singular N2, or with determiners that are not overtly marked for plural, cannot manifest number incongruence, and hence do not have an overt recognition mark of the postdeterminer status of the type noun. However, all type nouns used as postdeterminer are phonologically reduced (cf. Denison 2002: 3), which also reflects that they do not have autonomous head status.

The postdeterminer use of type nouns differs from the attributive modifier use in that, unlike the latter, no variant with *a* in front of N2 is possible. We find, for instance, as variant of *what sort of mother are you* in (31)

- (55) Whatever I did seemed to be wrong. I began to think “**what sort of a** mother am I to allow this to continue”. (www.truefreedomtrust.co.uk/testimonies/Abonusforlife.html/)

but none of the postdeterminer uses quoted above allow this variant:

- (43a) *In the North, **this sort of a** style was not at all popular.
 (44a) *It takes time to find **that sort of a** man, ...

Thus, in this section, we have argued that the postdeterminer use of type nouns constitutes a grammatical category in its own right, distinct from both head and modifier use, with its own functional meanings and with its own formal features and grammatical behaviour.

3.4 Qualifying uses

For the characterisation of the categories in the previous sections – head, modifier and postdeterminer use – a number of criticisms and amendments of the positions in the existing literature have been proposed. By contrast, the qualifying uses of type nouns with which this section is concerned have, in our opinion, been insightfully discussed in the literature. Denison (2002) offers a plausible scenario for the shifts leading to these uses and Aijmer (2002) gives a very fine-grained, data-based description of their various pragmatic values to which little can be added.

In their qualifying use, type noun strings basically hedge the ‘type specifications’ incorporated by the linguistic predications which they hold in their scope. In Quirk et al.’s (1972: 4552) terms, they function as ‘approximators’, ‘hedgers’ or ‘downtoners’ of these categorisations, as illustrated by

- (56) But I suppose it's as a that's as a **sort of** holiday, kind of doing you know nothing but sitting around. (*ICE-GB*, quoted in Denison 2002: 2)
- (57) Overall, *Reign of Fire* is a good popcorn film but not a spectacular film. I wont give anything away but it feels **sort of** idiotic and a waste of Matthew McConaughey's character.
(www.bigscreen.excepc.com/ReaderReview.php?movie=35783)
- (58) Now Stan Collymore had moved up to the A-Listers, having his way with Hollywood siren Sharon Stone in the back of a car. For 30 seconds. Well, **sort of**. (www.sky.com/showbiz/article/0,,50001-1178469,00.html)

Sort of in (56) indicates that the nominal category 'holiday' is used ironically, in (57) it downtones the predicate 'idiotic', and in (58) it modalises the whole preceding proposition, taking back some of the strength of that assertion. According to the basic type of grammatical unit being qualified, we will make a distinction between nominal, adverbial and sentential qualifiers, as illustrated by (56), (57) and (58) respectively.

In the nominal qualifier use, Denison (2002: 11) points out, the type noun has undergone a shift from head to modifier status. With reference to Langacker (1991), he (2002: 11) notes that the type noun is thus defocused and moves "towards a less holistic construal of [a] category, recognising its internal structure (degrees of centrality)". The qualifying construction is "in effect about the nature of membership of the class of N2" (*ibid.*), that is, it can be used to refer to a 'possible', 'arguable', or a 'peripheral' member of that class. Denison (2002: 11) also notes that *type*, which acquired the postdeterminer use much later than *sort* and *type*, does not take part in this development.

According to Denison (2002: 12), one of the scenarios by which the adverbial construction might have arisen is by extending the syntactic range of *sort of* or *kind of* as a hedging device from modifying a noun as in (56) to modifying other categories as in (59), in which *kind of* is used first with a verb, and then, after a pause to reframe the VP, before a noun.

- (59) and they **kind of** group – put people into **kind of** categories (*LLC*, quoted in Denison 2002: 12)

The extension of the qualifying use of *sort of* and *kind of* to whole propositions seems to represent a further step in this scope extension.

3.4.1 Nominal qualifier

In general, the nominal qualifying construction with *sort of* and *kind of* is, as pointed out by Denison (2002), concerned with the relation of its referent to the categorisation used in N2-position. The speaker can, for instance, be ironic about the categorisation (60) or show uncertainty about the category used (61).

- (60) They had allowed the economy to ride on a tide of credit and debt and called it an economic miracle. --- Well, it was a **kind of** miracle. After all, they made almost £100 billion on North Sea Oil revenue vanish without trace. (*CB – Today*)
- (61) He's a **sorta** Leftist (with whom I often find myself in agreement) whose blog is always worth a visit. (www.blimpish.typepad.com/blog/2005/04)

Qualifying *sort/kind of* are also often used as metalinguistic operators (Aijmer 2002: 178). As pointed out by Aijmer (2002: 195-196), they can function as marker of lexical imprecision, as a warning of a style shift, or as a self-conscious marker of creative idiom or metaphor, which "enables the speaker to be creative, to use words in an innovative and humorous way, to borrow phrases belonging to literary or more formal style and to use a slangy turn of speech" (Aijmer 2002: 195). For instance,

- (62) Well I don't like these er ... what's ... for want of a better word these **sort of** organised paths like the Pennine way. (*CB – UK spoken*)
- (63) Essentially, Richard Julin, the curator for the museum, was curious about the linkages between architecture, history, and how film can be seen a [sic] **kind of** hyper-textual archaeology. (www.djspooky.com/art.html)
- (64) Offering the viewer a **kind of** a la cart theme park ride through the future, "The Fifth Element" reminded me an awful lot of the Paul Verhoeven megamovie "Total Recall".
(www.dcox.customer.netspace.net.au/fifth.html/)

3.4.2 Adverbial qualifier

In their adverbial use, *sort of* and *kind of* can similarly be used as approximators to qualify the categorisations in their scope as approximate or imperfect in relation to the instances being depicted (Quirk et al. 1972: 452; Huddleston and Pullum 2002: 623-624; Bolinger 1972: 223; Aijmer 2002: 49; Margerie 2005). In some cases, they can also be used as downtoners, "suggesting an incomplete or low degree" of, for instance, the quality referred to (Quirk et al. 1972: 452). Types of grammatical unit that can be adverbially qualified in this way include adjectives, numbers, verbs and whole predications.

When used as an adverbial qualifier of an adjective as in (65), the type noun string indicates that the description is only approximate, or that the quality described is not present to a very high degree in the instance being considered.

- (65) Then later in the night we took a walk in our underwear around the campus. That was **sorta** weird.
(www.yaledailynews.com/article.asp?AID=23414)

As illustrated by (66), the qualifying type noun string can also serve as an indicator of number approximation (Aijmer 2002: 196):

- (66) it was the late ‘70s and everyone was getting into punk rock, and I was **kind of** sixteen and going for something else.
(www.npr.org/programs/thistle/features_reader_int2.html)

Likewise, with verbs, adverbial qualifiers can function either as approximators or as downtoners (Quirk et al. 1972: 276-454ff). In (67), the type noun string can be rephrased as ‘you could almost say’, which implies a denial of the truth-value of what is denoted by the verb. In (68), the type noun string invokes an assumed norm with reference to which the force of the verb is reduced.

- (67) He **sort of** smiled at us. (Quirk et al. 1972: 454)
- (68) Yeah I mean I’m not, I’m not saying you’re a bad influence on her now, I think she’s ... just **sort of** jumping in at the chance for someone to victimise personally [...] (*COLT*)

When the type noun string frames whole predications, it can occur in front of them or in end position, as in

- (69) This plugin is designed to solve that problem, **sort of**.
(www.nbcs.rutgers.edu/~hedrick/typography/index.html)

As noted by Quirk et al. (1972: 456-457), front position has to be distinguished further into pre-operator position, where the qualifier lies “within the scope of clause interrogation and negation” (70) and post-operator position, where it merely frames the type specifications of the predication (71).

- (70) He **kinda sorta** maybe loves him.
(www.alykat.hispeed.com/unfrozen/fanfic/mooks/index.htm)
- (71) I’ve **sort of** become part of the mountain bike world in a way without actually having a mountain bike; it’s quite strange.
(*COBUILD* – UK spoken)

3.4.3 Sentential qualifier

In its (as yet rare) uses as sentential qualifier, *sort of/kind of* have rid themselves even further of the structural constraints of the NP and VP. In these uses, they have scope over a whole proposition, and function like adverbs such as *perhaps*, *largely*, etc., qualifying the truth or accuracy of that proposition. Because of their propositional scope, they may involve conversational moves by distinct speakers, as in (72).

(72) CMG: Sort of like In The Fishtank?

Beam: **Kinda**. Definitely like that but not with all the para-meters. Not like you have to go and perform it all in one day. (www.cokemachineglow.com/feature/interview/beam.html)

3.5 Discourse marker use

As discourse markers, type noun strings are not tied to grammatical class boundaries anymore, and lack clear notions of scopal domain. They apply more diffusely to the discourse, through which they are scattered. According to Denison (2002: 4, 14), the discourse marker use developed from the adverbial use through processes of subjectification and semantic bleaching. As discourse markers, type noun strings are used as indicators of tentativity, and as fillers and hesitation markers. They may convey general tentativity about what is being said, comparable to the present-day discourse marker use of *like* (Meehan 1991), which is often used together with the type noun-string in this sense, as in (73). This is probably a further development of the 'approximator' value of *sort of* and *kind of*.

(73) As I remember it used to be **sort of** like fairly common for a Tuesday, that I'd pretend to be sick. (*ICE-GB*, quoted in Denison 2002: 4)

It can also be noted that tentative *sort of* also signals social non-dominance by the speaker, to the extent that it may be over-used by a speaker in an inferior position, as a marker of deference to the other speaker in superior position, as in the following exchange between lecturer (L) and student (S), quoted by Martin (1980).

(74) L: Do you find the system, er, makes sense?

S: Yes, I think I'm er beginning to understand it better now and I must say, I'm **sort of** impressed ...

I quite like the idea of, er, **sort of**, er, **sort of** flexibility I think is the key word, isn't it? ... in the **sort of** Prospectus.

L: So they tell us.

S: And yeah, and it seems to me as well that the, er, the exam system is, er, a much better idea if you **sort of**, er, take it, the whole, the whole, um, six terms and, um, you know, work it out on assessment like that rather than **sort of** three hours, pass or fail, **sort of**.

L: How are you getting on with Don Juan?

S: Er, um, I quite like it, really.

It's, er, part of it I find, er, **sort of**, er, a bit contrived but I suppose that's **sort of** Byron's style that, er, um

L: What is Byron's style?

S: One I think generally he's, er, it's, um, somewhat satirical style in that

L: Uh huh.

S: Uh, he's particularly in Don Juan he's **sort of** bringing out the, er, bitterness of **sort of** family life, I mean his wife, er, left him on, er...

Sort of and *kind of* may also be strewn across discourse as fillers or hesitation markers like *err*. Whereas some of these seem to be used largely unintentionally (75), other uses are more strategic (76), and serve the function of floor-holding and buying time “while the speaker plans what to say next” (Aijmer 2002: 188).

- (75) ...and I **sort of** opened the door, and looked out, and I **sort of** saw Richard... (Denison 2002: 4)
- (76) ^well I !don't think .^it's ^((**sort of** a)) . a complete con:cl\usion= you're **sort of** ^left with the - - you ^**sort of** [ɔ:m] – it's ^**sort** [ɔ:] an :end to a :story in a :w\ay= . you can ^just im'agine_these_things_going \on# it ^sort of !winds \up# it's [ɔ:m].^rather an _art!ficial .{^[d'u:nei'mal]}# ^rather 'like [?] 'one of [ɔ:m] 'Moli!ere's 'plays# (COLT, quoted in Aijmer 2002: 189)

3.6 Marker of onomatopoeia and quoted speech or thought

Finally, type noun strings can function to frame onomatopoeic expressions (77) and, like innovative quotative markers such as *go* and *be like*, represented speech (78).

- (77) I've ^never s\een a 'sortof# ^bottle 'after :b\ottle# . **sort of** ^pop 'pop p/opping# ãll the t\ime# - - and ^everybody got :awfully dr\unk I rem/ember# (COLT, quoted in Aijmer 2002: 186)
- (78) “I just got a visual, Sharon standing in front of the class going, (SCREAM), while these little kids **kinda** ‘Señorita Flynn? Hee hee hee hee hee.” (*Santa Barbara Corpus*, quoted in De Smedt 2005: 112)

In these uses, *kinda/kind of*, *sorta/sort of* are often used together with *be* and *go*, as in (79) and (80).

- (79) im just **being kinda hey** i can hear murkin
(www.livejournal.com/users/andyhello)
- (80) He **kinda** went, ‘Yeah, I think so!’
(www.rockconfidential.com/Testament.html)

3.7 Conclusion

In Section 3, we have argued that six functionally distinct categories of type noun uses have to be distinguished. In support of these, we have adduced semantic and formal criteria, many of which had been pointed out in descriptive grammars and specialised studies of type nouns. Particular attention has been paid to the delineation and elucidation of the NP-internal functions of type nouns – head, modifier and postdeterminer uses – which had been relatively neglected in the

previous literature. Yet, these NP-internal functions are not only important uses observed in synchronic data, they also constitute possible steps in the diachronic development of type noun constructions. In this respect, the subjectification and grammaticalisation processes observed by us in both the modifier and postdeterminer uses add to the full developmental picture, in which, before now, mainly the subjectification involved in qualifying and discourse marker uses had been stressed.

4. Corpus analysis of the functions of type nouns

In this section we will briefly report on how the descriptive framework proposed in Section 3 was applied to two sets of British English data. The first data set consists of exhaustive extractions on the type nouns *sort/kind/type + of* from *The Times*, a subcorpus of the *COBUILD* corpus. The second data set comprises exhaustive extractions on *sort/kind/type + of* (and *sorta/typalkinda*) from *COLT*, *The Bergen Corpus of London Teenage Language*. These two data sets (De Smedt 2005) represent the two ends of various register clines. The language in *The Times* is in the formal, written mode, and is, because of its normative function in British culture, not thought of as representing very progressive usage. The *COLT* corpus, by contrast, contains the highly informal, spontaneously spoken language of teenagers. According to Halliday (1978), casual conversation between peers constitutes the most important locus of language change and innovation, without being subject to overly conscious forms of monitoring or engineering.

Table 2 represents the relative frequencies of the six main functions of type nouns in *The Times* and *COLT* data sets. This quantified analysis relies strongly on De Smedt (2005).

Table 2: Corpus analysis of type noun functions in *The Times* and *COLT*.

	<i>The Times</i>		<i>COLT</i>	
	#	%	#	%
Head use	353	20.56	3	1.8
Modifier use	102	5.94	9	5.35
Attributive	74	4.3	3	1.78
Semi-suffix use	28	1.63	6	3.57
Postdeterminer use	923	53.75	16	9.52
Qualifying use	330	19.22	86	51.2
Nominal Qualifier	293	17.06	9	5.36
Adverbial qualifier	32	1.86	72	42.85
Sentential qualifier	5	0.29	5	2.97
Discourse marker use	7	0.41	40	23.8
Quotative use	1	0.06	6	3.57
Unclear	1	0.06	8	4.76
<i>Total</i>	1717	100	168	100

When we compare the quantified data from *The Times* and *COLT*, it is immediately clear that the relative frequencies peak for different clusters of functions in the two data sets. In *The Times* corpus the NP-internal type noun uses predominate, particularly the postdeterminer and head functions, and the nominal qualifier use. In *COLT*, on the other hand, it is precisely uses that are not tied to NP-structure that are most frequent, viz. adverbial qualifier and discourse marker use.

The large number of postdeterminer, head and nominal qualifier uses of type nouns in *The Times* and the prevalence of adverbial qualifiers and discourse markers in *COLT* can be explained by a variety of factors. Obviously, these differences correlate with the different stylistic properties and communicative aims of the two registers. The differences in the use of type nouns also reflect the different tendencies of the two registers with regard to degree of innovativeness. We will now discuss both these points in more detail.

The typical uses of type nouns in *The Times* data are directly or indirectly concerned with the classification of entities. In **head** noun function, which accounts for one fifth (20.56%) of all occurrences, the type nouns straightforwardly describe subclasses as being related to superordinate classes. This form of taxonomisation and subclassification is clearly rather important in newspaper texts.

The **nominal qualifying** function of type nouns is only slightly less frequent in *The Times* corpus (17.06%). The most common use made of this qualifying construction in *The Times* has the writer signalling to the reader not to take the expression in question literally, but to interpret it as an illustrative approximation or metaphoric reformulation of a discourse referent, as in (81).

- (81) Full breathing is also an important tool for encouraging waste elimination a **kind of** spring-cleaning process that can go on all year around, every day of your life. (*CB – The Times*)

By far the most frequent function of type nouns in *The Times* data, however, is the postdeterminer use, which constitutes more than half of all occurrences (53.75%). As we saw in Section 3, NPs with **postdeterminer** uses of type nouns realise anaphorically or cataphorically motivated reference to generalised concepts. These generalisations often express a property of the instance(s) ana- or cataphorically referred to, which is then applied to (an)other instance(s) as well. For instance, in (50) *these sort of scare tactics* refers back to ‘linking plastic milk bottles and cancer’ and at the same time refers to any other instances falling under that common denominator. In (52) *the kind of peace Stalin and Hitler brought to Poland* compares the peace installed by the Dayton accord with the peace brought to Poland by Stalin and Hitler. These and other examples quoted above illustrate the peculiar discourse effects which these NPs have, such as expressing a personal assessment of specific instances as flowing from generally accepted properties of other instances. The functional category of phorically motivated generalised reference masks, as it were, the fact that highly personal evaluations

are being expressed and sanctioned. De Smedt (2005: 119) also suggests that this form of reference with its abstract, objectified aura serves to draw the reader into this evaluation, and create an evaluative consensus between speaker and hearer, as illustrated by (82)

- (82) And when Brosnan swings his gun around a corner, it comes to a rock-steady halt. It's **the sort of** skill that comes in handy amid the swift, crisply defined action sequences of the film by far the best since *The Spy Who Loved Me*. (*CB – The Times*)

The high frequency of uses concerned with 'classification' of entities no doubt generally correlates with the high incidence of heavy NPs in *The Times* corpus. As noted by Halliday (1985b), NPs tend to do a lot of the meaning-making work in written texts such as quality journalism. Moreover, the way in which these NP-internal type noun constructions are used plays a role in the relation which is set up between writer and readers in this kind of journalism. They help the writer establish an authoritative tone, in which s/he projects expertise and sure judgement.

In the *COLT*-data, by contrast, we are dealing with a register of language use in which there are fewer heavy NPs, and in which VPs do a lot more work: spontaneously spoken language tends to juxtapose and combine shorter clauses with simpler NPs. The largest sets of type noun uses are formed by adverbial qualifiers (42.85%) and discourse marker uses (23.8%). The adverbial qualifiers mostly qualify adjectival and verbal phrases, that is, predications of utterances. Inspection of the data shows that in most cases the qualifiers play down the force of the utterance. The discourse markers are scattered throughout the discourse and convey tentativeness and solidarity. The high incidence of adverbial qualifiers and discourse markers in the teenage language recorded thus has the general effect of conveying non-dominance and of establishing solidarity.

As for the innovative character of type noun uses, it is striking that strong semantic shift and detachment from NP-structure is very much associated with the *COLT* data, and barely represented in *The Times*. The relative frequencies of the new uses in *COLT* may even to a certain extent reflect the progression of changes manifested by type noun constructions. Adverbial qualifiers, in which the type noun has shifted away from the qualification of entity-categorisations, are by far the most common in the *COLT*-data (42.85%). The further step in which the type noun string qualifies a whole proposition accounts for a much smaller portion (2.97%). The discourse marker use, hypothesised by Denison (2002) to be a further development of the adverbial qualifier, is well represented too (23.8%). By contrast, the newer quotative uses appear in only 3.57%. The only structurally innovative use attested with some frequency in *The Times* data is the semi-suffix use. *The Times* examples of semi-suffix use are still outnumbered by the attributive modifier data, despite the collocational restrictions that can be observed in the latter. It seems fairly uncontroversial to state that, as far as

frequent and creative semi-suffix uses are concerned, Internet uses are playing a more important pioneering role than the examples found in *The Times*.

5. Conclusion

In this contribution we have presented a functional classification of type noun uses, viz. six categories defined in terms of the main structural and semantic differences manifested in the data. In comparison with the categories proposed in the existing literature, we have advocated further distinctions mainly for the more congruent NP-internal uses, distinguishing modifier and postdeterminer uses from head uses. Following the literature on the subjectified uses of type nouns, nominal, adverbial and sentential qualifiers were distinguished from discourse and quotative markers. This descriptive framework was applied to two British English data sets from opposing registers: written texts from the quality newspaper *The Times* (*COBUILD* subcorpus) and spontaneously spoken conversation between teenagers (*COLT*). Quantification of these analyses has revealed strong asymmetries in the relative frequencies of the various type noun uses in both data sets. Type nouns in *The Times* were predominantly NP-internal and were concerned with the classification of entities in different form; these uses, it was suggested, contribute to the authoritative voice of the writer. In the *COLT*-data, adverbial qualifiers and discourse markers predominated; these convey tentativeness and solidarity. The figures thus overwhelmingly confirm the hypothesis that spontaneous conversation between peers is the prime locus of language change and innovation.

Notes

- 1 We are very grateful, for helpful suggestions and comments on earlier versions of our analysis of type noun constructions, to David Denison, Roberta Facchinetti, Stig Johansson, Tine Breban, Lieven Vandelanotte and Peter Willemse. We also thank David Denison for his permission to quote material from his paper "History of the *sort of* construction family", presented at the Second International Conference on Construction Grammar, Helsinki, 6-8 September 2002. We are also grateful to Keith Carlon for reading this article and suggesting unobtrusive but careful changes.
- 2 The examples from the *COBUILD* corpus quoted in this article were extracted via remote log-in and are reproduced with the kind permission of HarperCollins Publishers. All these examples are followed by the abbreviation *CB* as well as by the name of the subcorpus from which they were extracted. Internet examples are quoted with the url at which they were accessed.

- 3 Brems (2003, 2004) has investigated the shift from head to modifier found with measure nouns such as *load(s)*, *heap(s)*, and *pile(s)* in synchronic corpus data. She has interpreted this shift in terms of ongoing grammaticalisation.
- 4 The *COBUILD* data are not coded for intonation, so we are relying on our own reading of the examples here.
- 5 These non-official spelling variants are not systematic or consistent in informal registers in the case of pre-modified modifier uses. What's more, the *typa*-spelling occasionally gets extended to head uses in informal contexts, as in *What paper should I use? What kinda pencils and pens are best to draw with? Is it okay to copy pictures?* (www.organicmetal.co.uk/pages/tutorials_qa.htm/)
- 6 Even though the type noun is overwhelmingly singular in semi-suffix use, our data even contained one example in which the type noun is plural: *You might find that the people who tend to follow JFK kinds of theories are those who are highly motivated to look for external causes to a negative event* (CB – *The Times*). Hence, even the singular status of the type nouns can be stated only as a very strong tendency of semi-suffix use, not as an absolute rule.
- 7 We are using the notion of 'framing' here in the sense of McGregor (1997): a framing element indicates that what is framed is set apart from the 'first-order' discourse as in some sense meta-represented, i.e. quoted or reported.

References

- Aijmer, K. (2002), *English discourse particles: evidence from a corpus*. Studies in Corpus Linguistics X. Amsterdam: John Benjamins.
- Bache, C. (2000), *Essentials of mastering English. A concise grammar*. Berlin/New York: Mouton de Gruyter.
- Bolinger, D. (1972), *Degree words*. The Hague: Mouton de Gruyter.
- Breban, T. (2002), *Adjectives of comparison: postdeterminer, epithet and classifier uses*. Unpublished MA Thesis. Linguistics Department. University of Leuven.
- Breban, T. (forthcoming), 'Grammaticalisation and subjectification of the English adjectives of general comparison', in: B. Cornillie et al. (eds.) *Subjectivity and subjectification. Proceedings of the theme session "Paths of subjectivity" at ICLC8*. Berlin/New York: Mouton de Gruyter.
- Breban, T. and K. Davidse (2003), 'Adjectives of comparison: the grammaticalisation of their attribute uses into postdeterminer and classifier uses', *Folia Linguistica* 37 (3/4): 269-317.

- Brems, L. (2003), 'Measure noun constructions: an instance of semantically-driven grammaticalisation', *International journal of corpus linguistics*, 8: 283-312.
- Brems, L. (2004), 'Measure noun constructions: degrees of delexicalisation and grammaticalisation', in: K. Aijmer and B. Altenberg (eds.) *Advances in corpus linguistics: papers from the 23rd international conference on English language research on computerised corpora (ICAME 23)*. Amsterdam: Rodopi. 249-265.
- Denison, D. (2002), 'History of the *sort of* construction family'. Paper presented at the *Second international conference on construction grammar* (Helsinki, 6-8 September 2002).
- De Smedt, L. (2005), *Functions of the T-nouns kind, sort and type: a comprehensive, data-based description*. Unpublished MA Thesis. Linguistics Department. University of Leuven.
- Halliday, M. (1978), *Language as social semiotic. The social interpretation of language and meaning*. London: Arnold.
- Halliday, M. (1985a), *An introduction to functional grammar*. London: Arnold.
- Halliday, M. (1985b), *Spoken and written language*. Deakin: Deakin University Press.
- Halliday, M. (1994), *An introduction to functional grammar*. 2nd edition. London: Arnold.
- Hopper, P. (1991), 'On some principles of grammaticalisation', in: E. Traugott and B. Heine (eds.) *Approaches in grammaticalisation*, Volume 1. Amsterdam: John Benjamins. 17-36.
- Huddleston, R. and G. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Kruisinga, E. (1932), *A handbook of present-day English*. 5th edition. Groningen: Noordhoff.
- Langacker, R. (1991), *Foundations of cognitive grammar*. Volume 2: *Descriptive application*. Stanford: Stanford University Press.
- Lavrysen, J., E. Nys and E. Vreven (2005), 'Type noun constructions'. Term paper. Linguistics Department. University of Leuven.
- Lehmann, C. (1985), 'Grammaticalisation: synchronic variation and diachronic change', *Lingua e Stile*, 20: 303-318.
- Mackenzie, J. L. (1997), 'Grammar, discourse and knowledge: the use of *such*', in: J. Aarts, I. de Mönink and H. Wekker (eds.) *Studies in English language and teaching. In honour of Flor Aarts*. Amsterdam: Rodopi. 85-105.
- Margerie, H. (2005), 'On the semantic-pragmatic aspects of the grammaticalisation of *kind of* and *kinda* in British and American English'. Paper presented at the *FITIGRA Congress*, K.U. Leuven, 10-12 February 2005.
- Martin, J. (1980), *Text grammar*. Course notes. Linguistics Department. University of Sydney.

- Martin, J. (1992), *English text: systems and structures*. Amsterdam/Philadelphia: John Benjamins.
- McGregor, W. (1997), *Semiotic grammar*. Oxford: Clarendon.
- Meehan, T. (1991), 'It's like, "What's happening in the evolution of *like*?": a theory of grammaticalisation', *Kansas working papers in linguistics*, 16: 37-51.
- Oxford English Dictionary online*. Oxford: Oxford University Press. Available from: www.oed.com.
- Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1972), *A grammar of contemporary English*. London: Longman.
- Traugott, E. (1995), 'Subjectification in grammaticalisation', in: D. Stein and S. Wright (eds.) *Subjectivity and subjectivisation*. Cambridge: Cambridge University Press. 31-45.
- Traugott, E. (1996), 'The role of discourse markers in the theory of grammaticalisation'. Paper presented at the *ICHL*, Manchester 1995.
- Ward, G. and B. Birner (1995), 'Definiteness and the English existential', *Language*, 71: 722-742.
- Willemse, P. (2005), *Nominal reference point constructions: possessive and esphoric NPs in English*. Unpublished doctoral dissertation. Linguistics Department. University of Leuven.

This page intentionally left blank

Second Language Acquisition, parallel corpora and specialist corpora

This page intentionally left blank

Student writing of research articles in a foreign language: metacognition and corpora

Francesca Bianchi* and Roberto Pazzaglia**

* University of Lecce

** University of Pavia

Abstract

The aim of this paper is to describe the creation and use of an ESP corpus for teaching Italian undergraduate students how to write research articles in English. A metacognitive approach to reading-comprehension processes, integrated with corpus tools, was considered a prerequisite to writing activities. The students were guided in the analysis of the ways in which psychology experimental papers are structured. A corpus of psychology articles divided into moves was created (462,772 words; annotated and non-annotated parallel versions) and used in class to teach students how to analyse concordances for lexico-grammatical, and rhetorical reference in writing activities. Thanks to the availability of corpus concordances, even those students who had never used English when speaking or writing managed to produce scientific paragraphs which were grammatically and stylistically acceptable. Although the annotated files could not be used in class, the individual annotation activity carried out by the students had the primary pedagogic function of increasing awareness of the macro-structure and function-form relationships in experimental articles.

1. Introduction

The paper describes the creation and use of an ESP corpus in a module designed to teach Italian undergraduate students how to write research articles in English. Corpora have been and are constantly and fruitfully used in a wide variety of theoretical and applied settings, including describing general and specialised types of language, analysing genre, studying learners' interlanguage, compiling dictionaries and grammars, studying ideology and culture, studying translation issues, stylistics, forensic linguistics, teaching language, testing, just to mention a few (for a review of theoretical and applied uses, see: Hunston 2002; McEnery and Wilson 2001). More limited seems to be the use of corpora as writer's aids, and most of these uses are limited to terminology (Hunston 2002). However, some authors seem to have tentatively taken into consideration the idea that corpora may be convenient reference tools for writing. This argument is presented in Hunston (2002: 135) as follows:

For many writers who are expert in their own field [...] it is not the technical terminology but what might be called the terminology of rhetoric that causes problems. This is a problem facing, for example,

experts in academic disciplines attempting to write papers in a language not their own. For them, signals of organisation and purpose may be more difficult to use than the technical terminology. Corpus analysis of particular kinds of paper may be used to identify the words and phrases linked with specific moves and functions, the aim being to provide on-line assistance which will supply the writer with the most appropriate phraseology at each point of the article (Noguchi 2001). The disadvantage of this approach is that a large amount of research is needed to provide a resource for a relatively small group of people.

On a less theoretical basis, Bowker and Pearson (2002: Ch. 10) offer ideas of how to use a small corpus of articles from a scientific journal to highlight typical semantic and syntactic patterns in individual sections of scientific articles (Introduction and Discussion sections) and use them for writing. Their method is based on concordance lines of selected words from a non-tagged corpus.

The project described in this paper originated from preliminary considerations that are similar to Noguchi's ones and applied and expanded Bowker and Pearson's suggestions in order to create a teaching module for Italian students of psychology. The results suggest that, by taking advantage of a very contextualised setting and integrating corpus techniques into a metacognitive teaching approach to genre, home-made small/medium-sized dedicated corpora can prove extremely useful, for relatively little effort. The results of our study depended on the interaction of several theoretical and practical components: a highly standardised genre; specific ideas about writing; a metacognitive approach to teaching; and the high potential of corpora and corpus analysis tools. Each component is analysed separately in the following sections.

1.1 On the research article as a genre

Genre literature is extremely rich and varied, and several different 'schools' have dealt with the issue, including systemic linguistics, language anthropology, the ethnography of speaking, the sociology of language, the new rhetoric, and ESP. Each tradition and each author differs from the others in the way they focus on one or more of the following dimensions: structure, context, cognition, social structure, audience, and language (a clear and systematic review is provided by Paltridge 1997). With reference to the scientific article and for the purpose of teaching writing, it can be useful to categorise the studies into only two major macro-groups: those that match the surface elements of articles to historical and sociological contextual factors and foreground the dynamic processes that have brought the genre into being and that constantly re-shape it; and those that, without ignoring or denying the existence of contextual drives, foreground the linguistic features of the final product. The former, often arising from some kind of sociolinguistic perspective, tend to see a tight connection between form, content, and social dynamics, and focus on such elements as intertextuality, the creation of the author's status, and the author's need to find a suitable place

within the community. The latter describe the genre in terms of moves, obligatory vs. optional elements, lexico-grammatical features, and register.

It is our opinion that despite (or possibly owing to) the different traditions, most of these studies should not be considered as competing with each other; rather, like the tiles of a puzzle, they add new perspectives to the overall view of a genre. Studies foregrounding process are necessary to frame and conceptualise the activity of article writing and explain the existence of and the logic subtending structural and linguistic features.

The teaching project described in this article arises from a theoretical framework that is in keeping with the concept of 'family resemblance' (Swales 1990) and 'prototypicality' (Paltridge 1997; Swales 1990). As far as the development and purpose of scientific genres is concerned, the framework draws on different sociological traditions, while the perspective used in the analysis and teaching of genre features is very much in line with ESP approaches.

Far from being a simple question of grammar and vocabulary, the practice of research article writing is a complex activity grounded in a discipline's methodology and culture and embracing both content and form (Berkenkotter and Huckin 1995). Research articles tend to share conventional features that have been built up in the course of time by expert members of the scientific community (Bazerman 1988), on the basis of previous texts and textual elements (Berkenkotter and Huckin 1995: 17). Compliance with these conventions helps meet the audience's expectations (Günthner and Knoblauch 1995; Swales 1990), thus facilitating comprehension on part of other members of the same community, identifying the author as a member of the scientific community, and making the piece of writing more easily acceptable to the community (descriptions of peer review activity and hassles over the publication of scientific articles such as those described in Berkenkotter 1990, Berkenkotter and Huckin 1995, and Myers 1990 may provide interesting insights into this matter). At the same time conventions represent a convenient way to 'routinise' the writing of articles (Bergmann and Luckmann 1995; Günthner and Knoblauch 1995).

In terms of form, structural and rhetorical conventions can be observed in scientific articles (Swales 1990),¹ which are sometimes described in writing style manuals. Once learned, these conventions may simplify the practice of writing: in fact, structural conventions (that mirror content conventions) represent a well-trodden path for topic organisation and provide the author with a 'considerable repertoire of argumentative strategies' (Berkenkotter and Huckin 1995: 65).

Psychology research articles, like other types of research articles, tend to show a fairly limited number of recursive structural elements, often organised according to conventional textual schemata. A lengthy description of major structural and rhetorical features of this type of articles can be found in the *Publication manual of the American psychological association* (2001).² This, however, was not meant as a linguistic description, but rather as a series of suggestions from psychologists and editors to prospective authors. Therefore, for the purpose of the teaching module described in this paper, a preliminary analysis of a small number of psychology research articles randomly taken from different

specialised journals was carried out to identify typical moves and steps. The analysis, inspired by Gopnik's (1972), and Gläser's (1995) descriptions of research articles, as well as from the *Publication manual of the American psychological association*, highlighted nine moves, roughly corresponding to the main sections of an article (Abstract, Literature Review, Introduction, Method, Results, Discussion, Conclusion, Notes, Thanks). Of these, seven refer to the body of the article, i.e. are closely connected to the experiment and the way it is presented, while the remaining three (Abstract, Notes, and Thanks) can be considered as corollaries having different social purposes: the abstract is a 'selection tool' on the basis of which a psychologist decides to read the entire article or discards it as not in line with his/her research interests (Bazerman 1988); notes allow the introduction of information that for some reason the author wants to mark as 'secondary level'; finally, thanks allow the acknowledgement of help, support or suggestions of colleagues and authorities – a dutiful move in a restricted community.³ Finally, titles (including the main title, section titles, and captions) can be considered a sort of 'metamove' in that, taken together, they provide a summary of the structure/content of the article. Furthermore, from a purely linguistic point of view, they are all textual elements that may consist of an incomplete sentence. According to Bazerman (1988), summaries and titles go hand in hand with abstracts to help readers decide about the relevance of an article to one's own interests. Within each move, discourse progresses along a series of steps. The same steps tend to recur in several moves, with different levels of detail. The moves and steps used in the current project are summarised in Appendix 1. They were preliminarily selected by the teacher, presented to the students in class and used by the students in the creation of the corpus and for the interpretation of concordances.

On a different linguistic level, recurring lexico-grammatical, syntactic, and rhetorical patterns seem to be observable within and across moves and steps. Some of these patterns (such as prevalence of passive forms and nominalisation) are generally widespread in scientific or academic discourse, have been long analysed in the literature (see for example: Butler 1990; Flowerdew 2002; Gläser 1995; Gotti 2003; Halliday 1997; Huddleston 1971; Trimble 1985), and are well known to linguists and teachers alike. Others are discipline- and/or content-specific; hence the idea to have the students investigate recurring linguistic phenomena in a highly contextualised framework and in a significant number of texts, which can only be achieved through a specialised disciplinary-article corpus. The corpus should allow investigation within individual moves as well as in more than one move at a time.

1.2 On writing

Writing is a very complex type of activity and as such should be the focus of systematic training for both native and non-native speakers. In particular, writing a scientific paper requires competence at several levels: 1) general lexico-grammatical skills; 2) knowledge of technical and sub-technical vocabulary; 3) knowledge of how to convey meaning in a coherent and cohesive text; 4)

knowledge of the structural and rhetorical ‘habits’ that have been developing within the specific community. All of these competences are generally acquired through time and extensive contact with the language.

The need to comply with specific genre conventions and rhetorical habits is felt by some to be a limit on personal creativity, when not an unnecessary imposition of our society for the benefit of those who belong to a particular social group (Clark and Ivanič 1997: 53). In contrast, we believe that genre conventions and accepted rhetorical practices may be valuable resources for writers in general and non-native speakers in particular, as they represent a well defined guideline for individual writing in a particular field.

It is common practice for highly proficient non-native speakers involved in writing activities, such as academics, scientists, and translators, to keep to hand a text that is similar in content and form to the one they are writing and refer to it when they are stuck with a term or a way to express a concept. Interestingly, this is not instinctively done by writers with less experience or lower language proficiency, such as students, who tend to rely entirely on their own forces and (limited) competences. This frequently generates stress and leads to generally lower results.

The metacognitive-and-corpus approach to writing described here is an extension of this reference principle to a small but dedicated corpus of psychology research articles. The corpus, as well as being the basis for discovering general linguistic ‘trends’ of the genre, becomes a source of useful expressions, language chunks, and preferred sentence structures that the students can adopt and re-use in their own texts. It could be argued that, in its most extreme form, this type of writing activity becomes a patchwork of prefabricated chunks of language, where creativity is compromised. However, a series of counter arguments are put forward here. First of all, creativity is also apparent in the way pieces are put together to create a patchwork. Second, in the case of scientific papers, acceptability standards (generally shared by the scientific community) and journal review practice set limits to the scope of individual linguistic creativity; learning these limits should be part of the basic training for all young or would-be researchers. Finally, creativity can only be adequately employed by native or very proficient non-native speakers, who have full mastery of standard morphosyntactic and lexical features of the language used. For lower levels of proficiency, resorting to accepted phrases and structures avoids the risk of making gross errors and may help these students improve their general language abilities.

1.3 On the metacognitive approach

Metacognition refers to awareness of the existence of higher level cognitive processes, awareness of the choices such processes are guided by, and the ability to activate and control such processes. It plays a major role in providing a strategic attitude to problem solving activities and in applying and controlling the chosen strategies. Metacognitive theories have been applied to teaching in several ways, frequently focusing on specific cognitive abilities, such as mathematical

abilities, memory, and reading comprehension. The basic elements of all metacognitive approaches include knowing how the human mind works, being aware of one's own cognitive processes, being able to select and direct strategies to a goal, monitoring performance, and creating a positive and proactive image of oneself (Zanetti and Miazza 2004). Central to all applications of the metacognitive approach is the active participation of the subjects in the process of learning (Ashman and Conway 1991; Brown 1980; Garner 1987; Ines 1991, 1996). Knowledge is not acquired but rather co-constructed by the participants by means of conscious interaction with the task under investigation. Therefore, the role of the teacher is to guide the students in the acquisition of learning strategies and procedural knowledge (Garner 1987). An important factor that contributes to the achievement of results with metacognitive approaches is motivation (Borkowski, Carr, Rellinger and Presley 1990; De Beni and Moè 2000; Ines 1991, 1996). Motivation is connected to external factors, such as marking, gifts, bonuses, or fear of punishment, and internal factors, including the desire to learn, to be better than colleagues, and being convinced of the relevance of what one is doing.

Metacognitive theory has long been applied to reading comprehension tasks. Reading comprehension is similar to a problem-solving activity: the correct understanding of a text emerges only from the application of suitable cognitive schemas that include the abilities to draw inferences and to distinguish among the different parts of the text (Kintsch 1994). A good reader is aware of a comprehension aim in the reading activity, is able to recognise main content elements and relevant details, makes use of textual elements, such as titles, to lead comprehension, knows the existence of different types of text and the logical structure of each of them, is able to judge the text in terms of coherence, clarity, and complexity (Zanetti and Miazza 2004). Several metacognitive training programs in reading comprehension have been suggested. In particular, De Beni and Pazzaglia (1989, 1991) advance two different, but not necessarily incompatible, paths: one focuses on promoting lexical competence, activating semantic comprehension, acquiring or improving knowledge of textual features, controlling reading strategies and acquiring a flexible approach to texts; the other path aims to achieve awareness of a goal in reading and to promote strategic competence and control of one's own reading abilities. Finally, Cornoldi and Caponi (1991) and De Beni and Pazzaglia (1991) suggest focusing on a single genre and proceeding with explicit analysis of its structure and characteristics.

This project considered a metacognitive/metalinguistic approach to reading comprehension and genre analysis as a prerequisite to the writing tasks. This approach was integrated with corpus techniques to obtain enhanced results. In the module, the metacognitive approach was implemented at different levels. First of all, the entire module focused on a highly specific type of research article. Second, detailed analysis of major lexical, cohesive and other linguistic and textual features was carried out with the students (Part 1). Third, the students/would-be authors were made aware of the existence of genre conventions and of the advantages offered by complying with them and were

guided towards the analysis of the structure of experimental papers and the creation of a corpus (Part 2). At this level, the annotation activity was a means of promoting an active involvement of the students in the analysis of moves and steps and of mastering textual competences. Finally, corpus analysis techniques were offered to the students as new strategies for an autonomous investigation of texts, and control strategies were suggested for the writing process (Part 3). Interestingly, the centrality of the student in the learning process is a core feature in both metacognitive approaches to teaching and data-driven learning (corpus approach). This favours a seamless integration of the two types of approach.

1.4 On corpora

In this section we will try to explain how some basic resources of corpus linguistics, namely a home-made non-tagged corpus, concordance lines, clusters, wordlists, and keyword lists can be used as reference tools for LSP writing. We will try to do this by connecting corpus features to the theoretical and pedagogic framework outlined above. A theoretical and practical description of the corpus used in the project will also be made as a necessary introduction to its possible uses.

1.4.1 Corpus design

For the aims of the project, the most suitable resource was deemed to be a home-made corpus. This decision arose not only from the absence of a suitable corpus but also as part of the metacognitive approach. Selecting, reading, analysing, dividing, and annotating an article for the corpus were thought to be important learning moments for the students, as well as an active application of the theoretical knowledge acquired in class and a preparatory exercise for the use of the corpus for writing. However, corpus compilation always poses a series of issues in terms of content, size, and form. The following paragraphs describe the criteria adopted in the design of the *Psychology Corpus*.

In terms of content, the corpus should include only psychology experimental articles, possibly written by different authors and for different journals. The issue of whether authors should be all native speakers or speakers of one particular variety of English is irrelevant here. The corpus should be representative of the language of the psychology community, which includes authors from different nationalities using English as a *lingua franca*. The relevant point is that the articles in the corpus have been accepted for publication. In fact, published articles have gone through a long revision process in terms of content, terminological adequacy, grammaticality, and readability. When psychologists write an article or an abstract, their ultimate aim is to have their paper accepted for publication, and not to boast the most perfect native-speaker's competence. On the other hand, however, it could be reasonable to provide students with the best possible examples of language in use. Hence the decision in this project to select published articles having at least one native speaker among the authors,⁴ in

the hope that, as it frequently happens, s/he had made a linguistic revision of the draft and final copies of the paper.

In terms of size, several authors agree that the most specific the type of language to investigate, the smaller the corpus can be. On the basis of their experience, Bowker and Pearson (2002: 54) declared that

well-designed corpora that are anywhere from about ten thousand to several hundred of thousands of words in size have proved to be exceptionally useful in LSP studies.

Furthermore, some kind of compromise is always to be achieved between representativeness, manageability, and the time and resources available for the creation of the corpus. Therefore, we tried to narrow the scope of the corpus in terms of content by selecting groups of related keywords from the areas of language acquisition and developmental psychology (two major fields of interest for the students) and invited the students to use those keywords when searching for a suitable article to include in the corpus. A size of about 500,000 words was considered a reasonable compromise for the project.

A further issue to consider was annotation. This depends on the focus of investigation, but also on the availability of specific software tools. Lexico-grammatical features and rhetorical patterns can be easily retrieved even in a row corpus, starting from suitable search patterns and proceeding with a qualitative analysis of concordance lines (Bowker and Pearson 2002). POS tagging was briefly considered, as it would certainly represent a convenient type of annotation to simplify and automatise lexico-grammatical searches, but discarded for the time being, because no automatic annotation tool was available.

As for moves, the corpus was divided into sub-corpora/folders, each corresponding to a different move. This approach was in line with Gledhill's and Bowker and Pearson's approaches (Bowker and Pearson 2002; Gledhill 1995, 1996, 2000), with the major difference that their sub-corpora strictly corresponded to the different sections of articles (structural units) while in our case folders corresponded with moves (primary-level functional or communicative units). Moves may or may not overlap with the actual sections of articles. For example, the review of previous studies can appear as a separate section or be part of the Introduction, while concluding paragraphs can be integrated into the Discussion section rather than into the Conclusion section. To deal with such cases, the students were instructed to select the relevant parts of discourse (when they extended for a reasonable stretch of text) and move them into the most suitable folder. Such a subdivision of the corpus and detachment from the actual structure of the articles was functional to the main goal of the corpus, i.e. to provide a reference tool for writing in terms of preferred linguistic and rhetorical formulations. It was not our intention to create a corpus that could provide insight into the generic structure potential (Hasan 1989) of the psychology article as a genre. Possible moves were selected *a priori* on the basis of existing literature and the preliminary analysis of a limited number of

randomly selected articles; the sequential order of constituent parts in individual instantiations of the genre in question was felt to be dependent on content and author's personal choices.⁵

The last important element to take into consideration were steps, or secondary-level functional/communicative units. Although steps may at times be rather lengthy, they do not frequently stretch beyond a couple of sentences and in some moves, such as Abstract and Conclusion, can be even shorter. For this reason annotating the corpus seemed more practical than further subdividing it, and steps were annotated manually by the students.

Two parallel versions of the *Psychology Corpus* were thus created: an annotated and a non-annotated one. The corpus includes 67 empirical studies in psychology taken from 20 different international journals, with a total amount of 462,772 words. In terms of content, the texts cover the following topics: developmental psychology (31); general psychology (13); social psychology (14); psychotherapies (6); neuropsychology/physiology (3). All the articles were subdivided into moves and moves were grouped into folders. The following folders appear in the corpus: Titles; Notes; Thanks; Abstract; Introduction; Literature Review; Method; Results; Discussion; Conclusion. A spot review of the annotated version of the corpus, however, showed a considerable number of mistakes⁶ and suggested that it should not be used before a complete review of all the files was carried out by a linguist.

1.4.2 Corpus use

The annotated version of the corpus would allow us to carry out linguistic analysis at three different levels: entire corpus, individual moves, and individual steps. However even a version that provides only two direct levels of analysis (corpus and moves), as our non-annotated version, can be sufficient for the practical purpose of writing. Indeed, although the most sensitive level of analysis can theoretically be achieved with step-level investigation, some steps are nothing more than a lower-level repetition of existing moves, as Appendix 1 shows. This characterises move-level analysis as a good starting point for initial searches and the formulation of hypotheses about the rhetoric of articles, owing to a reasonable balance between size and specificity. Finally, a non-tagged corpus is easier to use with students who have no familiarity with corpus analysis tools.

With a suitable software tool, wordlists, keyword lists, concordance lines, and clusters can be easily produced from a row corpus. Each of these has its own role in the investigation of a corpus for writing purposes and may work as facilitator for the analysis of the results obtained with the other methods.

Sorted concordance lines may represent a direct view over meaning of individual words (sub-technical and general vocabulary being the most problematic for foreign language psychology students), phraseology, and lexico-grammatical structures (collocation and colligation). Each of these can easily be used by the students as starting points for the creation of correct and meaningful sentences. The very activity of sorting obliges the student to concentrate on typical English phrase and sentence structures, because a conscious decision must

be made to sort to the left or to the right. Comparison between concordance lines within and across moves may highlight preferred (more frequently used) rhetorical formulations.

Clusters help the students understand what to look for when reading concordances and may confirm the students' hypotheses about lexicogrammatical structures and phraseology. Furthermore, the fact that clusters are presented without context helps the student consider long chunks of language as re-usable structures for their own piece of writing.

In the *Psychology Corpus*, wordlists may be generated for the entire corpus or for individual moves. In either case, the wordlist may represent a general lexical resource for students, and a starting point for the selection of search words. When generated for a single move, it can also provide insight into the content of the move.

For the purposes of this project, useful keyword lists can be obtained comparing wordlists of different moves. The resulting lists highlight similarities and differences between moves/sections at several levels, such as content, and authorial stand. They may also help in the formulation of hypotheses about rhetorical habits of individual moves.

2. Project overview

This project was developed at the University of Pavia (Italy) within a 30-hour English language module in the degree course in psychology. The module, called *Scientific English*, focused on the experimental research article as a genre and had the following general purposes: analysing experimental research papers from a linguistic and genre perspective; teaching students new methods in learning, reading and writing; improving the students' reading and writing skills; creating a corpus of research articles in psychology that could be used by students, researchers and scholars of the Psychology Department as a language research tool and a reference tool for writing.

To achieve these goals, the module was divided into three task-oriented parts: Part 1 was designed to improve reading comprehension skills; Part 2 focused on the structure of research articles and the collective creation on the part of the students of a corpus of psychology papers divided into sub-sections; Part 3 was designed to teach students how to write research articles and abstracts. The students had 30 hours of classroom time, two hours per week, stretching over three terms in the academic year. Such a loose timetable was designed to give the students time to attend classes, review class content, do their homework, and take active part in the creation of the corpus. The students who attended lessons were allowed to subdivide the final exam into smaller, and therefore easier, parts (continuous assessment). Of the approximately 120 students enrolled, 78 attended lessons constantly and opted for the continuous assessment scheme. Continuous assessment tests took place at the end of each part. The following paragraphs

provide a detailed description of the *Scientific English* module, the objectives and organisation of each part, and examples of how the corpus was used.

2.1 Part 1: reading tasks

Part 1 was designed to improve reading comprehension skills. Since the students had had no contact with the English language for more than a year (i.e. since the end of the 30-hour *Basic English Language* module they are obliged to take in their first year), it was necessary to start with a general review of grammar. Subsequently, lessons moved on to the analysis of specific technical and sub-technical vocabulary and discussion of selected language functions, such as causality and result, analogy and contrast, and exemplification. Attention was also dedicated to analysing cohesive devices, English sentence and paragraph structures, and punctuation, as they noticeably differ from Italian. Starting points for theoretical explanations were selections from a course book for psychology students by Rossini Favretti and Bondi Paganelli (1988). Reading comprehension and grammar exercises were given as homework to the students in preparation for the first continuous assessment test which took place at the end of this first part of the course. Finally, at the end of Part 1, which coincided with the end of the first quarter of the year, the students were invited to individually select a paper that interested them from an on-line database of psychology research articles⁷ and were told that the article would be used in Part 2.

2.2 Part 2: analysis and annotation tasks

Part 2 focused on the structure of research articles and was designed to create an annotated corpus of psychology papers divided by move. The first lesson was dedicated to an introduction to corpora and their possible uses for language learning and research. The subsequent lessons were devoted to the analysis of the structure of the research article in terms of communicative and metacommunicative functions. In each lesson, the students were introduced to a different section of the research article and to a selected set of possible (or typical) moves and steps.

Analysis of selected paragraphs from psychology articles was then undertaken or shown in class as an example for the students. As homework, the students were asked to subdivide the psychology article they had chosen into moves and tag the content of each move using the functional and metacommunicative annotation system described in class (see Appendix 1). A copy of the original article, its non-annotated version and its annotated version had to be given to the teacher in electronic format before the beginning of Part 3.⁸ This task was a prerequisite for admission to the continuous assessment writing test (or to the final examination for the other students). At the end of Part 2, the second continuous assessment test was given; the test required students to analyse a few scientific paragraphs in terms of communicative steps. Before the start of Part 3, the students' articles were collected and the *Psychology Corpus* was created in its two parallel versions.

2.3 Part 3: writing tasks

Part 3 was designed to help the students write a research article or abstract using the corpus as their primary reference tool. Given the existence of creativity within the limits posed by acceptability conventions, a corpus approach helps highlight common semantic and grammatical trends across different individual texts belonging to the same family. Once noted, these ‘preferred’ lexicogrammatical formulations can be used by the students as starting points for the creation of correct and meaningful sentences.

Using the non-annotated version of the *Psychology Corpus*, the students were therefore taught how to produce and read concordances with WordSmith Tools (Scott 1996). Corpus concordances and other types of corpus queries were made available on a dedicated web page one or two days ahead of each lesson so that the students attending could print them out and take them to class; during lessons concordances were also projected on the wall for the benefit of those who had not managed to make printouts.⁹ To familiarise the students with concordance lines and the importance of co-text and context, attention was first dedicated to connectors, focusing on the ones that typically represent a problem for Italian native speakers (e.g. *however*; *nevertheless*; *because (-of)*; *due to*; *owing to*; *eventually*, and *finally*). Subsequently, the analysis moved towards some phraseological units, lexico-grammatical features and rhetorical features that recurred within and across the various sections of the research article. Concordance lines remained the main tool of analysis, but wordlists, keyword lists and clusters were also used when necessary. The students were asked to decide upon the meaning and use of each word on the basis of co-text and context and also to make generalisations or draw conclusions about verb tense usage and differences between sections/moves. They were then asked to apply what they had discovered and learned in quick guided writing exercises. Writing exercises that took longer to be completed, such as paraphrasing and summarising sections of research articles, or writing short paragraphs from notes, were assigned as homework. Occasionally, some grammar exercises were proposed to the students, focusing on aspects of the English language that tend to be problematic for Italian students, such as the use of determiners and modifiers, as well as word order within the noun phrase. At the very end of Part 3, the third continuous assessment test took place, where the students were asked to write an abstract in English from notes in Italian.

The following section provides examples of the type of concordances that were generated and analysed with the students, the generalisations that the students were helped to make, and the guided writing exercises they were presented with in preparation for the final writing task.

2.3.1 Examples of concordance analysis carried out in class with the students

Concordances were often generated for a single move, starting from words that might represent a typical or major step of that particular move. For example, in the Literature Review move, which focuses on past research, key words were considered words related to the concept of 'research', and authors' names. Concordances were generated for *study/studies*, *experiment/experiments*, *literature* and *research/researches* (the latter can be considered a false friend for Italian native speakers, as even very proficient students tend to use it as a countable noun to indicate one or more studies or experiments; the plural form was deliberately introduced by the teacher to show its absence). Authors' names were searched using the following pattern: (*) (Bowker and Pearson 2002). For the same search patterns, clusters were also taken into consideration. The students were then asked to compare the concordance lines and describe the use of each key word by completing a given table (Table 1). As well as being asked to take note of recurring phrases and verb tenses that accompanied the search words, the students were also led to draw some general conclusions about the typical use of those words and tenses in the particular move. Students were guided through a series of questions in such a way as to understand that when reference is made to research trends or previous literature in general (keywords *studies*, *experiments*, *research*, and *literature*) the present perfect tense is generally used, while the simple past is preferred to refer to a specific study or author. The simple present was also present as an alternative to the present perfect, but in a low number of hits; therefore, it was considered a stylistically possible variation adopted with limited frequency.

Verb tense usage was an object of discussion for almost all the concordance lines, in whatever type of move. Indeed, English scientific writing habits tend to differ from Italian ones in this respect and verb choice poses constant problems for Italian students writing in English. Concordances of words *aim/aims* and *goal/goals* in the Introduction section, for example, were used to show that the goal of the experiment is generally introduced in the simple past (e.g. *The aim of this research was to analyse*; *In this research we analysed*). Concordances and clusters of words *subjects* and *participants* in the Methods section were produced to see whether a trend exists in the description of experimental subjects and the administration of tests to subjects. The students were thus able to notice that the simple past is the most frequent tense for these purposes (e.g. *the subjects were 72 women*; *the subjects had 1 year of higher education*), and that both passive and active statements are frequently used in the description of test administration procedures (e.g. *subjects were asked*; *subjects were told*; *subjects were given*; *a group of subjects was chosen*; but also *we induced the subjects to*; *we asked the subjects to*; *the subjects wrote about*; *the subject knew*).

Table 1: Summary table of some key words used to describe previous studies.

<i>key words</i>	<i>phrases</i>	<i>rhetorical patterns</i>
Studies (plural!)	empirical studies experimental studies cross-sectional studies few studies	+ pres. perf.
Study (singular!)	present study current study this study (someone's) study a/an (adjective) study	+ simple pres. OR + simple past + simple past
Research (singular!)	recent research previous research considerable research research suggests that	+ pres. perf. OR + simple pres. OR + modal (may)
Literature	the literature on in the literature existing literature literature suggests that	+ pres. perf. + simple pres. + modal (can)

The students were puzzled by the fact that some sections show the same steps, yet they felt that some differences should exist. Comparison between wordlists of individual sections (keyword lists) provided some preliminary answers. Comparison of Abstract, Introduction, and Conclusion wordlists highlighted an interesting stylistic difference in terms of authorial 'presence': the author is almost completely invisible in abstracts (frequency counts: WE = 2.8%; OUR = 0%), but noticeably present in the Introduction section (WE = 62.8%; OUR = 31.11%), and only relatively visible in the Conclusion (WE = 39.28%; OUR = 28.57%). Finally, when observing Results and Discussion keyword lists, it was easy for the students to understand that while the Results section presents raw data (the keyword list was rich in terms referring to tables, graphs, and statistics), the Discussion section presents or discusses further analysis of the data (the keyword list showed preference for verbs in the present tense and a significant number of occurrences of the modal verb *may*, but no prominent semantic trend). The difference between the concepts of *results* and *discussion*, which sounds obvious to experienced scientific readers and writers, was not equally obvious to all students taking this module, as many of them had made mistakes in the splitting and annotating of their articles.

Although the types of analysis described in the previous paragraphs may sound a little superficial for linguists, they were sufficient for undergraduate students of psychology with limited English and very little experience as readers or writers of scientific articles, and proved satisfactory for the overall goals of the module.

At the end of each lesson, the students were asked to apply what they had discovered and learned by completing guided writing exercises. Cloze tests and paraphrase exercises obliged them to focus on lexico-grammatical, and rhetorical issues, while writing-from-notes exercises based on texts from the corpus represented the students' first steps in the creation of scientific paragraphs. This last type of exercise, which takes longer than the previous ones to complete and which was therefore assigned for homework, was preparatory to the writing part of the final exam (i.e. the third test in the continuous assessment scheme).

3. Students' results in the exams

Establishing the effectiveness of the described metacognitive-and-corpus approach in absolute terms is a very complex task. As is always the case with language teaching, results depend on the students as well as on the teacher. In this particular case a complete scientific assessment of the results of the teaching method should take into consideration all of the following parameters, at the very least: 1) students' final results; 2) students' proficiency levels; 3) individual learning abilities; 4) students' interest and motivation; 5) students' commitment in the reading, analysis, and writing tasks proposed; 6) students' actual use of the tools provided (concordance lines and notes); 7) active attendance at the course. Unfortunately, in our case, complete data were available only for parameters 1) and 7); the data referring to the other parameters were limited to a very few students. Hence complete statistical investigation could not be carried out.

The following paragraphs attempt a preliminary discussion of the pedagogic value of the module. This is based on the qualitative analysis of the writing tests of some students for whom information about use of concordances and commitment to the analysis and annotation tasks is known, followed by exploratory quantitative analyses based on the following parameters: students' results in the exam tasks, students' level of proficiency, and signed attendance sheets.¹⁰ Despite the limits of the quantitative and qualitative analyses presented below, they nevertheless provide interesting preliminary indications of the possibilities of a metacognitive-and-corpus approach.

Before proceeding with the discussion of the students' results, a brief description will be provided of the exam tasks that will be considered in the qualitative and/or quantitative analyses and of the assessment criteria adopted. In the analysis and annotation task, the students were requested to 'tag' the sentences of some scientific paragraphs according to the functional annotation scheme studied in class. The paragraphs were taken from psychology articles that had been read and used as examples during the course and the students were allowed to look at the annotation scheme (Appendix 1) during the exam. The writing task presented at the exam required the students to write a full abstract in English from notes in Italian. The notes referred to a local experiment to assess foreign language learning abilities, a general subject with which the students were familiar. On the exam explanation sheet relating to this task, general reference

was made to the ideal structure of an abstract (introductory sentences with reference to previous literature; description of the experiment; concluding sentences), and major assessment parameters (grammaticality and textual structure). As reference tools for writing the students were invited to use printed copies of concordance lines from the *Psychology Corpus* and class notes. The use of a dictionary was not allowed. When correcting the tests, the concept of grammaticality was interpreted in terms of sentence readability and compliance with the lexicogrammatical features observed in class. Basic grammar mistakes were not penalised unless they hindered comprehension. This assessment method was in line with the assessment method of some of the Cambridge ESOL examinations and with the types of exercise that had been carried out in class.¹¹ On the other hand, skipping a relevant move (such as not stating the goal of the experiment or not quoting previous studies) or presenting it in the wrong way or position was considered a penalising mistake, given the time dedicated to analysing the structure of abstracts in class. For all tests, marking was carried out in a blank fashion, to prevent results from being accidentally influenced by the teacher's knowledge of the student's names or proficiency in English. The marking range for all tasks was 0-8, sufficiency being reached with 5.

Appendix 2 shows three abstracts that were produced during the third continuous assessment test. They have been selected because the students who wrote them were known to the teacher by name and had been observed using or not using concordances during the exam. Student 1's abstract was marked 4.5 for the following reasons: no attention to the stylistic conventions highlighted in class (authors' absence in abstracts; preference for simple past tense in the description of the goal, subjects, material, procedure, and findings of the experiment); careless punctuation; incomprehensible penultimate sentence; failure to respect S-V-O word order in the subordinate clause of the last sentence. In terms of textual structure, on the other hand, the abstract shows no major faults.

Student 2's abstract was marked 6.5 (a 'fully sufficient' mark). In fact, the general structure is rather good, the lack of the 'Background' move being compensated by an explicit reference to literature when illustrating the experimental results; compliance to verb tense preferences is inconsistent but not absent; despite a few basic grammar mistakes (especially in the last sentence), readability is good and comprehension is never at risk.

Finally, Student 3's abstract was marked 7. Despite its extreme conciseness, it includes all the necessary moves, stylistic preferences are respected, and readability levels are very high. A glance at these students' proficiency levels, active attendance at the course, results in the analysis and annotation exam task, and use of corpus tools during the exam suggested possible explanations for these results. Student 1 was an advanced student (proficiency test score = 48) who actively participated in Part 2 lessons, which is mirrored in the high mark obtained in the analysis task (mark = 7); her participation in Part 3, however, was extremely inconsistent, when not disturbing, as she kept on chatting and laughing with friends during the lessons; at the exam she used none of the reference tools provided. Student 2 was a beginner (proficiency test score =

23.5) who attended all the lessons attentively, obtained 7 in the analysis task, and made ample use of concordance lines and notes during the exam. Finally, Student 3 was an intermediate student (proficiency test score = 35.5) who, like Student 2, attended all the lessons with interest and made use of concordance lines and notes at the exam. Her analysis task received top marks (8). In the three cases described, results in the writing task would seem to be directly connected to participation in the lessons, commitment in the analysis task, and use of the corpus tools provided rather than proficiency level.

As part of a search for general trends, quantitative analyses were also carried out on the final writing task results of all those attending and non-attending students who took or completed the exam before June 2004. Table 2 shows a comparison between attending and non-attending students (N = 78 and N = 34 respectively).

Table 2: Writing test results of the students who completed the exam by June 2004.

	<i>attending students</i>	<i>non-attending students</i>
fail	48.7	76.5
pass	51.3	23.5

In the group attending lessons, the failure rate was drastically lower than in the other group (40.7% against 76.5%), indicating that attendance was generally useful. Given that most (N = 62) of the students who attended the lessons had taken an assessment test at the beginning of the course, it was possible to examine results according to the students' levels of proficiency in English (Table 3).

Table 3: Writing test results of attending students, divided according to proficiency level.

	<i>fail</i>	<i>pass</i>	<i>pass marks</i>	
	<i>total</i>	<i>total</i>	<i>just</i>	<i>fully sufficient</i>
beginner	61.9	38.1	14.3	23.8
intermediate	59.3	40.7	29.6	11.1
advanced	21.5	78.5	7.1	71.4

The group included 21 beginners, 27 intermediate, and 14 advanced students. In a writing task of the kind proposed at the exam (writing a scientific abstract in English from notes in Italian), it could be reasonably expected that results would depend on proficiency level. The results in Table 3, however, show a fairly different picture.

As expected, the highest pass rate was obtained by advanced-level students (78.5%), and almost all the advanced students who passed this test wrote 'fully sufficient' abstracts. However, almost no difference is observable between the beginner and intermediate groups in terms of pass rate (38.1% vs. 40.7%), and, even more strikingly, beginners passed the writing test with higher marks

than intermediate students (23.8% of ‘fully sufficient’ marks for beginners against a mere 11.1% for intermediate students). This also means that some beginners fared much better than some intermediate and advanced students. In other words, results in the writing test were not directly related to proficiency, as Figure 1 clearly shows. On the other hand, a relationship between results in the writing test and the analysis and annotation test can be seen (Figure 2), although it is not of a direct type.

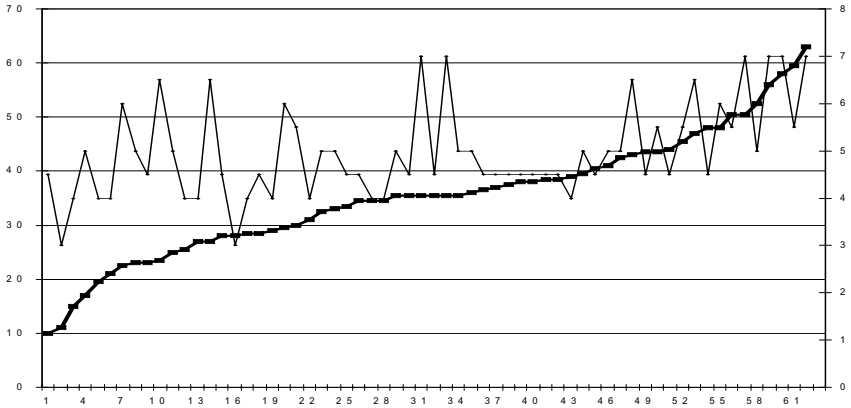


Figure 1: Writing test results and proficiency level: no relation exists between these two parameters.

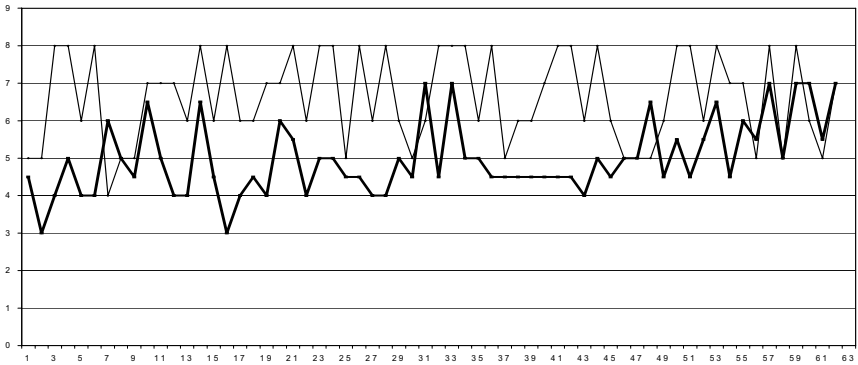


Figure 2: Results for writing test and analysis-and-annotation test: the two lines indicate the existence of a relation of an indirect type.

4. Discussion of the pedagogic aspects of the module

From a pedagogic perspective, the planning of this module stemmed from the belief that guided analysis of a limited number of research articles might be sufficient to familiarise the students with macro-structural elements and provide them with a valid skeleton framework for the creation of an individual paper or abstract, and that guided corpus analysis of collocations and grammatical patterns within and across structural elements could provide a valid tool when it comes to filling the ‘skeleton’ with content. The module linked a metacognitive approach, focusing on reading-comprehension abilities and genre analysis, with basic corpus tools and techniques, with the purpose of helping undergraduate psychology students to write correct scientific paragraphs in English. The metacognitive approach related to all parts of the module and acquired metalinguistic connotations in Parts 1 and 3.

From the students’ questions and comments during the first lessons, there emerged a general perplexity towards this approach: beginners thought it was too difficult for them, most intermediate and advanced students considered it too ‘unusual’ to be taken seriously, while a few advanced students were happy to do something different for a change. However, as the course proceeded, the students gradually started to see the point and benefits of the course. At the end of Part 2, some beginner students declared (unprompted) that the course had so far been useful not only as regards improving their general knowledge of English, but also as regards their ability to understand a scientific text whether in English or Italian as well as to memorise the content of scientific articles. The students attributed these ‘unexpected’ results to the annotation activity.¹² The annotation activity, however, as well as proving useful for the reasons noted above, was essential in making the students aware of the corpus they were contributing to, as well as of the discourse structure of the research article. Finally, the subdivision of the corpus into sub-corpora greatly facilitated the exploration of linguistic features of the genre under analysis. Finally, the creation of the corpus was taken very seriously by nearly all participants, but annotating a full article represented a long and difficult activity for these students who were not language specialists, and results did not always reflect the effort that the students had put into their work.

However, this realisation of the metacognitive-and-corpus teaching method was not free of limitations due to the small number of hours and the high number of students in the module. Time did not allow for a sufficiently smooth and gradual introduction to corpus analysis techniques and this probably increased some students’ resistance to this novel approach. Furthermore, the fact that it was not possible for the students to have hands-on access to the corpus and make their own enquiries certainly curtailed their curiosity. An ideal situation would also see the tags decided and agreed upon by the students themselves, rather than imposed on them. However this is clearly feasible only with small groups of students and many hours of classroom time. Finally, students of all levels could certainly have benefited from a higher number of general grammar

exercises, which would have probably improved the grammatical level of all the writing tests.

5. Conclusion

The results obtained support the use of corpus tools in teaching students of a scientific faculty how to write scientific paragraphs in English. By means of corpus concordances, motivated students who had never used English for speaking or writing managed to produce grammatically and rhetorically acceptable scientific paragraphs. However, from this experience with Italian psychology students, we would suggest that a corpus alone is not enough; the students need to be made aware of the existence of conventions and of the advantages offered by complying with them, to be guided in the discovery of the moves and steps that characterise scientific papers, and to be helped in understanding the rhetorical conventions that are used to realise such steps.

A metacognitive approach to reading comprehension and genre features can be easily integrated with a home-made specialised corpus to obtain the desired results. In the metacognitive-and-corpus approach described in this paper, preliminary reading and analysis phases focusing on experimental articles paved the way for a more autonomous and reasoned use of corpus queries. Individual annotation activity of one or more of the files in the corpus played an essential pedagogic role in increasing the students' awareness of textual structures and function-form relationships in research articles. It also provided the basis for the correct analysis and use of corpus concordances for writing purposes. In the writing activity, the use of a specialised electronic corpus divided by move allowed direct access to several texts at a time, in a non linear fashion. Non-linearity obliged the student to pay greater attention to language rather than content. Corpus concordances, clusters, wordlists, and keyword lists all contributed to the investigation of the corpus for writing purposes and each of them worked as facilitator for the analysis of the results obtained with the other query methods. Furthermore, as well as representing an essential element with respect to writing activities, the metacognitive approach also facilitated comprehension and memorisation of the content of scientific articles in English and Italian. Finally, it seems probable that the positive results obtained with this approach were connected to the focus in all the tasks performed (reading, comprehending, analysing, writing) on a genre with a wide range of standard characteristics.

Notes

- 1 Berkenkotter and Huckin (1995) emphasise the importance of content-related elements (such as 'appropriateness', 'novelty', and 'rhetorical timing') as intrinsically important features of genre writing, rather than

- form. The relevance of these elements and their impact on the various textual features is not questioned here, but given the type of students and the purpose of the course (psychology undergraduate students learning a foreign language) greater attention had to be dedicated to 'surface level' elements (language and structure).
- 2 The *Publication manual* is intended as a guide to authors and provides indications regarding content, structure and form. For a discussion of its history and development see Bazerman (1988: Ch. 9).
 - 3 This move is often performed in a separate section called Acknowledgments. The term *Thanks* was preferred here simply because it was easier for the students to remember.
 - 4 This was judged on the basis of name and affiliation.
 - 5 Corpus-based attempts to determine the generic structure potential and/or typical patterning of structural elements have at times been fruitless. See for example Paltridge (1997: 66-71) for an analysis of *Introductions* in scientific reports, and Gesuato (2001: 388-391) for an analysis of job application letters.
 - 6 Mistakes are a common problem in corpora created by students. For an analysis of advantages and disadvantages of learners' involvement in the construction of corpora, see Aston (2002).
 - 7 The database used was *OVID* (<http://gateway.ovid.com>), an on-line resource for which the University of Pavia has a subscription.
 - 8 Annotating a full article is a long and complex task, but the students were able to take advantage of a 3-week Christmas break.
 - 9 This type of approach was necessary because of the high number of students attending lessons (78) and the fact that it was not possible to have the students work in smaller groups without drastically reducing the number of hours for each group.
 - 10 Students' proficiency levels were measured at the very beginning of the course by means of a general multiple-choice grammar test that had been developed and used for years as a placement test in a local private foreign languages school with good results, and of a general vocabulary test developed by Palladino and Bianchi to test lexical abilities in adult learners of English (Palladino and Bianchi, in preparation).
 - 11 Information about the Cambridge ESOL examinations are available at the following address: <http://www.cambridgeesol.org/index.htm>.
 - 12 These results, that seemed 'unexpected' to the students because they extended to Italian, are some of the expected effects of metacognition, which is not a language-specific phenomenon.

References

- Ashman, F. A. and R. N. F. Conway (1991), *Guida alla didattica metacognitiva*. Trento: Erikson.
- Aston, G. (2002), 'The learner as corpus designer', in: B. Kettemann and G. Marko (eds.) *Teaching and learning by doing corpus linguistics*. Amsterdam: Rodopi. 9-25.
- Bazerman, C. (1988), *Shaping written knowledge. The genre and activity of the experimental article in science*. London: The University of Wisconsin Press.
- Bergmann, J. R. and T. Luckmann (1995), 'Reconstructive genres of everyday communication', in: U. Quasthoff (ed.) *Aspects of oral communication*. Berlin: Mouton de Gruyter. 289-304.
- Berkenkotter, C. (1990), 'Evolution of a scholarly forum: reader, 1977-1988', in: D. Roen and G. Kirsch (eds.) *A sense of audience in written communication*. Newbury Park (CA): Sage. 191-215.
- Berkenkotter, C. and T. N. Huckin (1995), *Genre knowledge in disciplinary communication: cognition/culture/power*. Hillsdale (NJ): Lawrence Erlbaum Associates.
- Borkowski, J., M. Carr, E. Rellinger and M. Presley (1990), 'Self-regulated cognition: interdependence of metacognition, attributions, and self-esteem', in: B. Jones and L. Idol (eds.) *Dimensions of thinking and cognitive instruction*. Hillsdale (NJ): Lawrence Erlbaum Associates. 53-92.
- Bowker, L. and J. Pearson (2002), *Working with specialized language. A practical guide to using corpora*. London/New York: Routledge.
- Brown, A. L. (1980), 'Metacognitive development and reading', in: R. J. Spiro, B. Bruce and W. F. Brewer (eds.) *Theoretical issues in reading and comprehension*. Hillsdale (NJ): Lawrence Erlbaum Associates. 453-481.
- Butler, C. S. (1990), 'Qualifications in science: modal meanings in scientific texts', in: W. Nash (ed.) *The writing scholar. Studies in academic discourse*. Newbury Park/London/New Delhi: Sage Publications. 137-170.
- Clark, R. and R. Ivanič (1997), *The politics of writing*. London/New York: Routledge.
- Cornoldi, C. and B. Caponi (1991), *Memoria e metacognizione*. Trento: Erickson.
- De Beni, R. and A. Moè (2000), *Motivazione e apprendimento*. Bologna: Il Mulino.
- De Beni, R. and F. Pazzaglia (1989), 'Nuove prospettive nella promozione della comprensione della lettura. Gli aspetti metacognitivi nel progetto MT', *Scuola e didattica*, 8: 24-26.
- De Beni, R. and F. Pazzaglia (1991), *Lettura e metacognizione*. Trento: Erickson.
- Flowerdew, J. (ed.) (2002), *Academic discourse, applied linguistics and language studies*. Hong Kong: Pearson Education.
- Garner, R. (1987), *Metacognition and reading comprehension*. Norwood (NJ): Ablex.

- Gesuato, S. (2001), *Job application letters as requests: characterization of a genre within speech act theory*. Bergamo: MG Editori.
- Gläser, R. (1995), *Linguistic features and genre profiles of scientific English*. Frankfurt-am-Main: Peter Lang GmbH.
- Gledhill, C. (1995), 'Collocation and genre analysis: the phraseology of grammatical items in cancer research abstracts and articles', *Zeitschrift für Anglistik und Amerikanistik*, 43: 11-29.
- Gledhill, C. (1996), 'Science as collocation: phraseology in cancer research articles', in: S. Botley, J. Glass, T. McEnery and A. Wilson (eds.) *Proceedings of the teaching and language corpora 1996*. Lancaster: UCREL. 108-126.
- Gledhill, C. (2000), 'The discourse function of collocation in research article introductions', *English for Specific Purposes*, 19 (2): 115-135.
- Gopnik, M. (1972), *Linguistic structures in scientific texts*. The Hague: Mouton de Gruyter.
- Gotti, M. (2003), *Specialized discourse: linguistic features and changing conventions*. Bern: Peter Lang.
- Günthner, S. and H. Knoblauch (1995), 'Culturally patterned speaking practices – The analysis of communicative genres', *Pragmatics*, 5 (1): 1-32.
- Halliday, M. A. K. (1997), 'On the grammar of scientific English', in: C. Taylor Torsello (a cura di) *Grammatica: studi interlinguistici*. Padova: Unipress. 21-38.
- Hasan, R. (1989), 'The structure of a text', in: M. A. K. Halliday and R. Hasan (eds.) *Language, text and context*. Oxford: Oxford University Press. 52-69.
- Huddleston, R. D. (1971), *The sentence in written English. A syntactic study based on an analysis of scientific texts*. Cambridge: Cambridge University Press.
- Hunston, S. (2002), *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Ianes, D. (1991), 'Introduzione all'edizione italiana', in: A. F. Ashman and R. N. F. Conway, *Guida alla didattica metacognitiva per le difficoltà di apprendimento*. Trento: Erickson. 7-36.
- Ianes, D. (a cura di) (1996), *Metacognizione e insegnamento*. Trento: Erikson.
- Kintsch, W. (1994), 'The psychology of discourse processing', in: M. A. Gernsbacher (ed.) *Handbook of Psycholinguistic*. Washington DC: Academic Press. 721-739.
- McEnery, T. and A. Wilson (2001), *Corpus linguistics*. 2nd edition. Edinburgh: Edinburgh University Press.
- Myers, G. (1990), *Writing biology: texts in the social construction of science*. Madison (WI): University of Wisconsin Press.
- Noguchi, J. (2001), *The science review article: an opportune genre in the construction of science*. Unpublished PhD thesis, University of Birmingham.

- Palladino P. and F. Bianchi (in preparation), ‘Improving foreign language learning at undergraduate level: Native language predictive variables and foreign language sensitive skills’.
- Paltridge, B. (1997), *Genre, frames and writing in research settings*. Amsterdam/Philadelphia: John Benjamins.
- Publication manual of the American psychological association* (2001). Washington D.C.: American Psychological Association.
- Rossini Favretti, R. and M. Bondi Paganelli (1988), *Il testo psicologico. Aspetti della traduzione e della lettura in lingua inglese*. Bologna: Pitagora Editrice.
- Scott, M. (1996), *WordSmith Tools*. Oxford: Oxford University Press.
- Swales, J. (1990), *Genre analysis. English in academic research settings*. Cambridge: Cambridge University Press.
- Trimble, L. (1985), *English for science and technology*. Cambridge: Cambridge University Press.
- Zanetti, M. A. and D. Miazza (2004), *La comprensione del testo. Modelli e ricerche in psicologia*. Roma: Carocci.

Appendix A

Titles

Title	<Tart> </Tart>	paper’s title
Subtitle	<SubTart> </SubTart>	paper’s subtitle
Section title	<Tsec> </Tsec>	section title or subtitle
Caption	<cap> </cap>	Caption

Notes

Author’s addresses	<mailto> </mailto>	details provided to get in contact with the author(s)
Notes	<txtnote> </txtnote>	notes to the text
Author’s details	<aboutauthor> </aboutauthor>	information about the author(s) and their role in the research project
Miscellaneous	<misc> </misc>	other type of notes

Thanks

Acknowledgement to individual	<indiv> </indiv>	thanking an individual
Acknowledgement to institution	<instit> </instit>	thanking an institution
Generic acknowledgement	<genak> </genak>	thanking generic groups of people
Grant	<grant> </grant>	mentioning or detailing a grant for the project
Stand	<stand> </stand>	author's stand with respect to funding or employing institution
Request	<req> </req>	request for comments, suggestions, etc.
Copyright	<copyr> </copyr>	copyright

Abstract/Introduction

Object	<obj> </obj>	aim of the study
Background	<backg> </backg>	reporting past research
Rationale	<rationale></rationale>	when reference is made to a theory
Subjects	<subj> </subj>	description of subjects
Material	<mat> </mat>	description of material
Procedure	<proc> </proc>	procedure
Method	<method> </method>	in case descriptions of subjects, material and procedure are not distinct
Findings	<findings> </findings>	results (expected or obtained)
Conclusions	<conc></conc>	final remarks – conclusions

Method

Subjects	<subj> </subj>	description of subjects
Material	<mat> </mat>	description of material
Procedure	<proc> </proc>	description of procedure

Literature review

Object	<obj></obj>	object or aim of previous study
Findings	<findings></findings>	results of previous study
Subjects	<subj></subj>	description of previous study's participants
Material	<mat></mat>	description of previous study's material
Procedure	<proc></proc>	description of previous study's procedure
Method	<method></method>	when descriptions of subjects, material and procedure are not distinct
Relevance	<relevance></relevance>	why are these previous studies relevant to the current experiment?
Discussion	<discussion></discussion>	author's comments about previous studies' methods, findings
Rationale	<rationale></rationale>	Rationale for the current study; to be used when reference is made to a theory and not to empirical past research
General statement	<gen></gen>	author's general considerations
Need for extension	<further></further>	need for further data, analysis, etc.
Limits	<limits></limits>	limits of previous experiments

Results/Discussion

Object	<obj> </obj>	aim of the study
Background	<backg> </backg>	reporting past research
Rationale	<rationale></rationale>	when reference is made to a theory
Subjects	<subj> </subj>	description of subjects
Material	<mat> </mat>	description of material
Procedure	<proc> </proc>	procedure
Method	<method> </method>	in case descriptions of subjects, material and

		procedure are not distinct
Findings	<findings> </findings>	results (expected or obtained)
Conclusions	<conc></conc>	final remarks - conclusions
Discussion	<discussion></discussion>	author's comments about subjects, material, procedure, etc
General statement	<gen></gen>	author's general considerations
Need for extension	<further></further>	need for further data, analysis, etc.
Limits	<limits></limits>	limits of current experiment

Conclusion

Apply suitable tags from other sections.

Appendix B

In this study we want to examine correlation between attitude of learning of the second language (L2) and the comprehension of the mother language (L1)

This study is supported by Correll and Armstrong's studies like: "Second Language Abilities" (1958) and "Foreign Language Learning Ability" (1962) where Working Memory and the comprehension of L1 are underlined in the learning of L2.

Subjects of this study are 136 psychology students of the 1st, 2nd, and 3rd year.

This research is based on a Pre-Test and Post-Test made by three tests where the first one is: Vocabulary, the second one is Grammar and the third one is Reading and Comprehension, that are administered at the beginning of the English course and at the end.

The hypothesis is controlled by analysis parameters like Course frequency, Grammar, Vocabulary and Comprehensions and by statistical analysis with SPSS.

The results show that are necessary other studies to prove the research hypothesis.

Student 1.

ABSTRACT

THE PURPOSE OF THIS STUDY WAS TO INVESTIGATE THE RELATIONSHIP BETWEEN L1 AND L2. THE SUBJECTS WERE 136 STUDENTS OF THE COURSE OF PSYCHOLOGY.

THE PRETEST IS CHARACTERIZED BY 3 TESTS FOR L1 AND 3 FOR L2. THE PRETEST AND THE POSTTEST WERE THE SAME ONLY IN L1: A GRAMMAR TEST, SOME VOCABULARY TESTS, READING AND COMPREHENSION.

THE PRETEST IS ADMINISTERED AT THE BEGINNING OF THE ENGLISH LANGUAGE COURSE, THE POST TEST IS ADMINISTERED AT THE END OF THE ENGLISH LANGUAGE COURSE. THE STATISTICS ANALYSIS ARE EXAMINED WITH SPSS. THE RESULTS WERE ON LINE WITH THE LITERATURE, BUT IS NECESSARY TO STUDY ANOTHER TESTS TO HAVE MORE INFORMATION ABOUT THE RELATIONSHIP BETWEEN L1 AND L2.

Student 2.

This study examined relations between L1 and L2.

~~Carroll~~ Relations between L1 and L2 had received theoretical attention in different ways. Carroll examined second language abilities, ~~Carroll~~ ~~Carroll~~ Pimsleur analyzed foreign language abilities.

Participants included 136 ^{psychology University} students of the 1st, 2nd and 3rd years.

Subjects were randomly assigned to each experimental condition.

The experimental session consisted of ~~two~~ ^{2 phases}; pretest and post-test. The pre-test consisted in ~~questions~~ ^{questions} in L1 and L2, the post-test only in L1.

Students were examined at the begin^{and at the end} of the ~~years~~ English course.

Students were assessed in English grammar, vocabulary and comprehension.

Findings from this research and from literature (Carroll, J.B, 1938; Carroll J.B and Sapon, 1933; Pimsleur, 1962) indicated ~~that~~ ^{that} other tests were necessary to ~~show~~ show L1 and L2 relation.

Student 3.

This page intentionally left blank

The structure of corpora in SLA research

Ron Cowan* and Michael Leiser**

* University of Illinois at Urbana Champaign

** Florida State University

Abstract

The field of Second Language Acquisition currently uses L2 error corpora to supplement its primary methodology – grammatical judgment and production tests – for investigating specific hypotheses about the development of interlanguages. This paper argues that large L2 corpora structured according to specific criteria would allow researchers to investigate with greater precision the contribution of the native language in the evolution of interlanguages as well as concepts such as overgeneralisation and end-state grammars. Furthermore, hypotheses posed about interlanguage development after native-like attainment, such as so-called U-shaped development, could be verified. The considerations for building L2 corpora of the future are illustrated by considering the validity and reliability of a multi-level corpora of errors produced by Spanish and Korean speakers learning English.

1. Introduction

At the 25th ICAME conference a number of provocative suggestions regarding the direction that corpus linguistics might take over the next 25 years were proposed at the panel discussion “I CAME, but where are we going?” Granger suggested that corpus linguistics could play an important role in Second Language Acquisition (SLA) research. This paper seeks to elaborate on this theme by demonstrating how large-scale corpora of adult second language (L2) learner errors could help answer specific research questions in this field.

There are three related areas of current SLA research where data from L2 learner corpora could advance our general knowledge of the development of Interlanguage (IL) grammars. The first is the investigation of the nature of ‘stabilised’ IL grammars. The second is the validation of U-shaped development in IL grammars that supposedly occurs after learners have attained native speaker competence in certain domains. Finally, L2 learner corpora could shed light on how the native language (L1) influences the learner’s progress through the different stages that have been posited for IL development. We will discuss these in the following sections.

1.1 Investigating the nature of ‘stabilised’ IL grammars

The focus on testing whether L2 learners use negative evidence and whether relationships defined by UG theory (such as the subset/superset relationship) affect L2 acquisition has abated somewhat due to the overall change in the

theoretical model proposed by Chomsky in the Minimalist program. Researchers are currently moving toward developing a more detailed picture of IL grammars and the extent to which they approach native-like competence. Studies like Birdsong and Molis (2001) strongly suggest that adults who begin learning an L2 as late as their middle twenties, well after the so-called critical period, can gain excellent, and indeed sometimes native-like grammatical proficiency in a second language. But it is generally accepted that most L2 learners eventually reach a point where their IL grammar stabilises – that is, in many respects it is native-like, but in some areas predictable spoken and written errors persist. The nature of such ‘stabilised’ (Long 2003) or ‘end-state’ grammars (White 2002) is an issue of considerable interest to SLA researchers today.

Attempts to study the nature of stabilised grammars must inevitably take into consideration the extent to which errors found in the later stages of the learner’s IL are attributable to the L1. Researchers with quite diverse theoretical viewpoints have gradually come to acknowledge the contribution of the L1, and in some theoretical models, such as Schwartz and Sprouse’s (1996) Full Transfer/Full Access model, transfer is assigned a significant role in the development of ILs.

Recently three researchers, Han (2000), Lardiere (1998a, 1998b, 2000a, 2000b) and Long (2003) have published reports investigating stabilised grammars. The method employed in all three is to collect error examples of an IL grammar from a few well-chosen learners. Han’s study is representative. He examined the extent to which a typical error made by Chinese learners of English, the so-called ‘pseudo-passive’ shown in (1), is found in the writing of two adult learners of English who had been living in the United States for three years, had scored over 610 of the TOEFL ten years ago, and had published articles in international journals.

(1) *The letter about graphics file has not received.

Han’s corpus was a 2-year collection of writing samples supplemented by a 10-sentence translation task and a grammatical judgment and correction task. This data allowed him to reanalyse the source of the error as the transfer of an L1 surface structure pattern and to conclude that very advanced learners would probably regress and continue to produce pseudo-passive constructions.

Although Han’s approach is certainly viable for studying stabilised grammars, it focuses only on one kind of error, and the validity and reliability of such a corpus may seem questionable to some researchers, due to its small sample size. Furthermore, this approach does not facilitate cross-linguistic comparisons that would reveal whether some error types are found in other languages that are typologically similar or different. To achieve these research goals and to provide a more complete developmental picture of the evolution of ILs, would require a much larger L2 error corpus that could be checked against experimental studies.

1.2 Validating the ‘U-Shaped’ Development Hypothesis

Related to the issue of the issue of stabilised IL grammars is the so-called ‘U-shaped’ development referred to in recent discussions about native-like attainment by L2 learners (Doughty 2003; McLaughlin and Hereda 1996; Segalowitz 2003; Norris and Ortega 2003). U-shaped behavior is a well-known phenomenon noted in L1 acquisition – the early accurate acquisition of certain linguistic forms is followed by a period of ‘back-sliding’, which is characterised by deviant forms that gradually disappear. Ervin’s (1964) study of English-speaking children’s acquisition of the past tense is frequently cited as an example of this U-shaped development. She noted that, initially, children acquire the correct past tense forms of irregular verbs like *go* and *break*. This is followed by a stage where *went* and *broke* are changed into the erroneous regular past tense forms **goed* and **brokeed*. However, eventually L1 learners return to the correct irregular forms. The so-called Uniqueness Principle, proposed in various versions by Roeper (1981) and Pinker (1984), offers an explanation for this U-shaped developmental sequence. This principle states that a given meaning in a language can have only one form. Children follow this in the acquisition of the past tense in English. When they first hear correct irregular past tense forms, they assimilate them. But as they encounter many more regular verbs in parental input, they are led to revise their initial hypothesis, and assume that *-ed* is in fact the correct representation of past tense in English. They adopt this revision and apply *-ed* to the irregular verbs in their lexicon. During this stage of their development, the two forms *broke* and **brokeed* may coexist, but as the input from adult speech continues, they gradually realise the past tense has two forms depending on the verb. So the provisional ungrammatical forms for the irregular verbs are replaced by the initially correct forms that they had produced earlier.

Kellerman (1985) was the first to propose that this phenomenon might also be found in SLA. He cites three cases in support of this, all of which involve Dutch learners of English and German. The first case describes a study of the acquisition of the verb *break* by EFL Dutch secondary school and university students. In Dutch and English *break* is a paired ergative verb, that is, it is both transitive (e.g. *He broke the vase*) as well as ergative (e.g. *The vase broke*). Kellerman and his colleagues found that up to age 17 the students had little difficulty in accepting both the transitive and ergative uses of *break*. But the following year, the last in high school, the students began to accept only the transitive use of the verb. This increased until age 20, at which time they began to accept transitive and ergative uses. The second case concerns English conditional sentences. Dutch has structures that correspond to English past habitual structures (*If he would come, we would go out*) and hypothetical conditional structures (*If he came we would go out*). Kellerman discovered that Dutch EFL students produced both of these up until the last years of high school but then began to favor a single structure that had *would* in both clauses. This could lead to ungrammatical hypothetical conditionals such as **If you would go to Chicago you would see a lot of beautiful buildings*. This stage persisted through the first year of college, and

for some learners Kellerman claims that it “is never eradicated” (Kellerman 1985: 350). The final case concerns Dutch speakers learning German idioms. Second-year university students tended to reject Dutch-like idiomatic expressions in German regardless of their grammaticality. Third-year students showed an ability to distinguish the idioms that were possible in German and those that were possible in Dutch. Kellerman concludes by noting that the “exact nature of the mechanisms responsible for these changes must for the moment remain a mystery” (Kellerman 1985: 353).

The U-shaped development in SLA described by Kellerman is counterintuitive. Most SLA theories envision a linear progression consisting of a series of stages where there is heavy reliance on the L1 at first, followed by a period of decreasing L1 influence where the learner works out the system of rules that may eventually be identical to the L2. Imperfect acquisition of specific structures such as the example cited above by Han, are assumed to be due to residual effects of the L1. Lately, however, the importance of determining the extent to which U-shaped behavior may occur in SLA has been revived in the context of investigating the concept of ‘fossilisation’ and stabilised IL grammars. Long (2003) notes that determining whether IL grammars are indeed stabilised requires research that can demonstrate that the appearance of competence in a certain subdomain of the L2 is not in fact illusory and subject to further development. Norris and Ortega (2003) also note that the issue of determining an accurate measure of when learners can be said to be using L2 target forms regularly is complicated by the issue of U-shaped behavior. Wolfe-Quintero, Inagaki and Kim (1998) have suggested the existence of an omega-shaped behaviour whereby L2 learners produce many instances of a new form, some of which are not accurate, followed by a drop in the instances of the form once it has been worked out.

There are a number of questions about U-shaped development in L2 acquisition that remain to be answered. Is it a typical characteristic of IL development that corresponds to what has been noted for L1 acquisition, or is it a temporary phenomenon that arises when “attentional capacity is depleted” due to “stress or informational overload” as Segalowitz (2003: 397) has conjectured? If it turns out that the U-shaped development is a regular developmental phenomenon, is it limited to lexical phenomena like verb classes or is it common to broader syntactic phenomena? These issues can be addressed more systematically if we possess structured L2 error corpora that show the performance of larger groups of L2 learners in different languages as their ILs progress through different stages of development.

1.3 Examining developmental sequences

A third area where L2 corpora may help expand our knowledge of IL development relates to the fixed stages that L2 learners are said to pass through as they move toward competence in the target language. Stages have been identified in the acquisition of negation, questions and word order, but recent research by Kathleen Bardovi-Harlig (1998, 1999) and her colleagues (Bardovi-Harlig and

Bergström 1996; Bardovi-Harlig and Reynolds 1995) indicates that L2 learners' acquisition of tense and aspect may also follow this kind of path. The so-called Aspect Hypothesis (Anderson 1991; Anderson and Shirai 1994, 1996) predicts that past and perfect forms will be first associated with telic verbs. The progression predicted by the Aspect Hypothesis is that perfective and past tense form first appear with achievement verbs, then proceed to activity verbs, and then to statives. Progressive or imperfect action will be associated with atelic forms, i.e. verbs that have no endpoints. Stative verbs will remain unmarked (have no endings added to them) longer than the other verbs. The acquisition of progressive forms will be applied first to stative and activity verbs and then spread to achievement verbs.

Evidence from the Bardovi-Harlig and Reynolds study, which involved 182 ESL learners representing 15 different languages stratified in six proficiency levels from beginner to advanced, found English achievement and accomplishment verbs to be the best carriers of past tense at all levels of proficiency. Stative verbs and activity verbs were less likely to have past tense applied to them in contexts where this is obligatory in English. Bardovi-Harlig and Reynolds concluded that their results support the Aspect Hypothesis and that ESL learners will acquire past tense developmentally in terms of different stages. First, telic verbs show a higher use of the past tense than state and activity verbs. Next, stative verbs show a higher use of past tense than activity verbs, and finally activity verbs gain parity with stative verbs. Initially there was a greater use of progressive forms with activity verbs and simple present forms with stative verbs before the simple past tense forms finally begin to predominate. A subsequent study by Collins (2002) with 136 French university students in Canada that used the same methodology as Bardovi-Harlig and Reynolds supports the claim that, over all, the progress of the learning of tense is shaped by the Aspect Hypothesis. But Collins found two important differences that did not turn up in the Bardovi-Harlig and Reynolds study. First, stative verbs were quite difficult for her ESL students. They rarely supplied past endings. She attributed this to the fact that Bardovi-Harlig and Reynolds did include enough statives. A second, more striking difference was that her subjects showed a different order of acquisition. The French-speaking students supplied present perfect forms for telic verbs in contexts where the simple past was obligatory. The most likely explanation for this is that in French the tense that would be used in these contexts is the *passé composé*, which in form resembles the English present perfect. The fact that Collins' results conflict with the Bardovi-Harlig and Reynolds' results indicates that the L1 also influences ESL students attempts to master English tense. Collins concluded, correctly in our opinion, that the interaction of L1 influence, lexical aspect and developing proficiency will be an important topic for future SLA research.

Although the experiments described above had large sample sizes, they did not always contain enough tokens to pick up trends, as the Bardovi-Harlig and Reynolds study shows. Furthermore, these kinds of experiments only sample the behaviour of subjects with specific levels of L2 proficiency. This limits their

ability to provide a developmental picture of how tense is acquired and the extent to which the L1 is implicated at different stages of development. Evidence from appropriately structured L2 corpora could play an important part in confirming the Aspect Hypothesis and revealing the interaction of the L1 and inherent lexical aspect in the acquisition tense and aspect in IL grammars.

2. Tracking errors in L2 corpora

To investigate the issues described above, it is important to consider the characteristics that L2 learner corpora should have. By way of addressing this we would like to examine what conclusions can be drawn about the persistence of L1 transfer errors made by native speakers of Spanish based on a corpus of written English. The purpose of this study is to identify and track written English errors of L1 Spanish learners. Specifically, the study investigates the following questions: which errors remain as a feature of their IL as they progress towards stabilised grammars? Which errors disappear as a result of continued exposure to the target language and instruction?

2.1 Subjects

The L2 learners who produced the corpus of errors were all native speakers of Spanish. They were enrolled in courses at the University of Illinois, all of which were designed to meet the needs of ESL students with different levels of proficiency. The group with the lowest proficiency comprises students in the Intensive English Institute (IEI), which prepares students to achieve a high enough TOEFL score to enter an American university. There are three levels in the IEI based on an institutional TOEFL test placement score. These range from low intermediate (Level 3) to advanced (Level 1). The TOEFL scores for this group range from 468 to 523. Students in ESL 114 and 115 have attained a TOEFL score that allowed them to be admitted as undergraduates. Their mean TOEFL score is 550. ESL 401 is the highest course required for international graduate students. TOEFL scores for this group are usually above 560. Because the Standard Error of Measurement for the TOEFL score is 14.1 for 95% confidence, it is clear that we have at least two and possibly three levels of English proficiency represented in the corpus, although there is an overlap between the undergraduate and graduate students.

2.2 Data

The corpus data produced by the students consist of first (unedited) drafts of written English. Each student produced three writing samples per semester. The word count breakdown by level was the following: IEI (7458 words); ESL 114/115 (12,143 words); ESL 401 (15,571 words). So it is not a particularly large corpus.

2.3 Error types

Ten distinctively different errors that are unambiguously attributable to L1 influence were selected to be followed through the different levels of proficiency in the corpus. These are shown in sentences (2) through (12) below. Designations in bold type after each sentence indicate the class (IEI, ESL 114/115, and ESL 401) from which the composition containing the error came.

The first error type is the omission of dummy *it* in English extraposition structures. This is due to the fact that, as a null subject language, Spanish does not contain an equivalent pleonastic pronoun. A corresponding sentence in Spanish begins with *es*. This results in the omission error shown in (2), where the \emptyset symbol marks the missing *it*.

- (2) * \emptyset Is very hard to live without privacy even in their home. **IEI 2**

The second error type is the insertion of definite articles before titles, e.g. (3a), names and abstract nouns, e.g. (3b), which is obligatory in Spanish but prohibited in English.

- (3) a. *They simple ignore and live *the life* with hope and faith **IEI 2**
 b. **The President Lyndon Johnson* undertook this policy in 1964. **ESL 115**

The third error is the free variation substitution of *another* and *other* before nouns. Spanish has only one lexical item *otro*, which has the meaning of *another* and *other* when they denote the two meanings “different” and “additional”.

- (4) a. *The virtual reality is *other important item* about research labs. **IEI 2**
 b. **Another professionals* in the civil engineering programs are trying. **ESL 114**

English adds the comparative morpheme *-er* to two-syllable adjectives ending in *-y*. Spanish forms the comparative by placing *más* (more) before all adjectives. This results in errors like those shown in (5). The fact that these errors occur frequently in other Romance languages, like Portuguese, and differ from the so-called “doubly marked” comparatives like *more easier*, *more nicer* (Biber et al. 1999: 525) suggests that they are indeed L1 induced.

- (5) a. *Second, I think is *more easy* to learn. **IEI 2**
 b. *The children who go with their fathers are *more happy*. **IEI2**

Spanish uses reflexive verbs to express the meaning of many English regular verbs and verbal idioms. This is carried over into English in cases like (6) where the writer intended to express the idea “make the effort” (*esforzarse* in Spanish).

- (6) *In that sense, if you want a great education you will have *to effort yourself* to learn more every time. **ESL 114**

Perhaps one of the most widely discussed errors is the transfer of the argument structure of L1 verbs whose meanings are equivalent to English modal verbs. Because the corresponding Spanish verbs take infinitival complements, errors like (7) are very common.

- (7) *In all situations I hear English and I *must to talk* in English, and the most important thing is that I *must to pronounce* correct English. **IEI2**

Nouns and adjectives in Spanish that are followed by specific prepositions like *para* and *de* take infinitival complements. The lexical equivalents in English take obligatory gerundive complements. In example (8) we see an example where the infinitive complement that would follow *es capaz de* has been carried over in English instead of “capable of passing”.

- (8) *The student who failed a test in the past and now he *is capable to pass* the test after some preparation knows that he is not a fool. **ESL 401**

Many phrasal and prepositional verbs in English do not require prepositions before objects in Spanish. This results in omission of the necessary prepositions when Spanish speakers attempt to express the same meaning in English, as shown in (9). The equivalent of the English prepositional verb *listen to* in Spanish is *escuchar*. It is not followed by a preposition, so (9) would be: *todo los dias escucho la radio*.

- (9) *Every day I *listen \emptyset the radio*. **IEI 2**

Two types of problems occur with Spanish speakers' rendition of English questions. Because there is no equivalent of a *do* support rule in Spanish, learners of English may omit the appropriate form of *do* in the formation of English *wh*-questions as illustrated in (10a). But it is also possible that (10a) is a stage 3 developmental error of the type described by Pienemann, Johnston and Brindley (1988). Unlike English, Spanish embedded questions involve both *wh*-movement and subject-aux inversion. Carrying over the L1 pattern into English results in errors like (10b).

- (10) a. **Why that happened* so many times in a short period? **ESL 401**
 b. *the question is *how can be imposed guidelines* to restrict [...] **ESL 114**

Finally, negation in Spanish is accomplished by insertion of the negative word *no*. There is no accompanying *do* support as in English, resulting in omissions of the type shown in (11).

- (11) a. *[...] do that if *he had not* his Ph.D. **ESL 114**
- b. *The facts show *there is not correlation* between the death. **ESL 401**

Table 1 shows the ten error types tracked over all course levels. There is no uniform pattern for all of the errors. Only three categories show a steady increase in errors across all levels – errors with definite articles, the substitution of infinitives for gerunds, and errors with prepositional and phrasal verbs. These must be examples of persistent errors, because they all have more errors at the highest proficiency level.

Table 1: Errors recorded across proficiency levels.

Error category		Proficiency Level					Total
		IEI 3	IEI 2	IEI 1	ESL 114/115	ESL 401	
1	deletion of dummy <i>it</i>	3	5	0	1	6	15
2	articles	3	17	7	13	34*	74
3	<i>another/ other</i>	1	1	1	4	4	11
4	comparative form of adj	1	1	1	0	0	3
5	reflexive verbs	0	0	0	1	1	2
6	modals	1	3	1	0	0	5
7	infinitive for gerund	1	11*	2	7*	18	39
8	prep verbs	3	7	1	13	22	46
9	questions	0	3	2	2	8	15
10	negation	0	6	0	3	9	8
							218

*indicates that a single writer made at least 5 errors in this category

Categories 2 and 7 should be persistent because they represent a learning problem that combines syntax with lexical learning. As the students’ proficiency increases, they will use new vocabulary and they may not master the appropriate prepositions that are obligatory with certain verbs and adjectives. The explanation for these errors is thus only partially L1 interference, whereas the problem of incorrect article use is solely L1 transfer. Two error categories – 4, comparative forms of adjectives and 6, following modals with infinitives – are both mastered perfectly before the learners reach university level. Although there is some slight increase with category 3, confusion with *another* and *other*, there are so few errors that we would expect that this would disappear. Category 5, the creation of English reflexive verbs based on semantically equivalent reflexive verbs in Spanish is virtually non-existent in this corpus. Categories 1, 9, and 10 show some slight increase in errors at the 401 level, but these increases are accompanied by interesting changes. For example, the dummy *it* in English extraposition structures no longer is deleted at the beginning of the sentence. Instead, this only occurs within embedded clauses as shown in (12). Presumably

this would be the last stage before total disappearance; however, it might also be a feature of a stabilised IL.

(12) *To answer that, \emptyset is important to make a clear distinction between [...]

Negation errors are generally of the type shown in (13), indicating that the learners are still having trouble working out cases where *no* is the only choice. And finally, errors with questions are essentially limited to working out the correct linear order of embedded questions, e.g. (14). This suggests that these three errors are about to disappear permanently in the English of Spanish speakers.

(13) *[...] an example of agnosticism *is the no acceptance* of confession.

(14) *I was afraid about *what did the persons asked me*.

Also, a number of new error types shown begin to appear in the writing of high proficiency learners. The well-documented phenomenon of pro-drop (omitting subject pronouns) that occurs in Spanish and other Romance languages is not carried over into English in lower level compositions, but it now appears in complex sentences, like (15).

(15) *When a child grows up \emptyset *tends* to imitate all that he sees.

In addition, the insertion of frequency adverbs between verbs and objects, permitted in Spanish but not in English, begins to occur more frequently, producing errors like (16a) and (16b). Quantifier errors that reflect almost word-for-word renditions of Spanish, e.g. *todo el público* for (16c) are now found. The pronoun *it* begins to be used for S anaphora where the demonstrative pronouns *this* or *that* are required in English, e.g. (16d).

- (16) a. *and they *miss a lot* their family
 b. *the tests *have also* their disadvantages
 c. **all the audience* clapped furiously
 d. *As beneficial as plant regulators can be for increasing growth, *it* represents a manipulation of nature.

3. Discussion

The overall picture presented by this corpus is one of some errors rapidly disappearing, some persisting into the advanced level with no sign of diminishing, others showing a slight increase, and new errors emerging as a result of the pressure to write more complex structures in the L2. One error, the insertion of definite articles before titles, names and abstract nouns appears to be a strong candidate for having the status of a persistent error in Spanish speakers'

stabilised IL grammar. The evidence is not quite as compelling for prepositions triggering the use of incorrect complements and the incorrect rendering of English prepositional verbs.

There is no evidence of a U-shaped development in this corpus, but one interesting phenomenon is the shift of the extraposition errors to a new environment as learners attain advanced level proficiency. Norris and Ortega (2003) note that the extension of a grammatical form into other contexts is an unexplored phenomenon of L2 production. The data suggest that this is a one aspect SLA research where L2 corpora can make a significant contribution. Recall that Kellerman’s data indicated that U-shaped development is found in the acquisition of lexical phenomena like ergative verbs. There were no instances of errors with ergative verbs in the Spanish data. This may be due to the fact that ergativity is more restricted in Spanish (Zobl 1989) and also to the overall size of the corpus. If we track the development of ergative verbs in a much larger corpus, we might better be able to find traces of this phenomenon. Table 2 below shows the total number of incorrect passivisations per occurrence of ergative verbs made in compositions of Korean students enrolled in the same courses as the Spanish students. A concordance search carried out on 30 English ergative verbs, most of which take the passive morpheme in Korean, found a very high rate of overpassivisation errors of the type: **It is ridiculous that women should be suffered from this kind of abuse.* (This type of error is discussed in greater detail in Cowan, Choi and Kim [2003].)

Table 2: Overpassivisation of English ergative verbs by Koreans.

Verb	ESL 114/115		ESL 401	
	%	# Tokens	%	# Tokens
change	30	10	62.9	27
consist	0	2	0	7
continue	16.7	6	38.8	18
decrease	12.5	8	3.5	16
disappear	25	4	40	5
exist	66.7	3	25	4
happen	0	8	10.5	19
improve	50	2	66.7	12
increase	0	12	25	36
last	25	4	0	6
occur	18	11	19.23	26
originate	0	0	20	5
result	0	0	20	5
suffer	0	11	25	12
vanish	0	0	12.5	8

(From Cowan et al. 2003)

The combined size of the undergraduate (ESL 114/115) corpus of errors, 60,236 words, and the graduate (ESL 401) corpus, 115,988 words, provides a reasonable

level of confidence for concluding that the overpassivisation of English ergative verbs is indeed a feature of the stabilised grammar of Korean learners of English. So it would appear that U-shaped development found for some phenomena like ergative verbs may not be present with some L2 learners. The Korean data also suggest that relatedness between the learner's L1 and the L2 may be a factor in the presence or absence of U-shaped development in IL grammars.

The data in the two corpora above allow us to draw some conclusions about the structure of L2 learner corpora for SLA research. The corpus of Spanish speaking learners of English is clearly too small to make any definitive claims about the development of ILs. At best, it raises interesting questions that could begin to be answered if we increased it five-fold to the size of the Koran corpus. It would also be desirable to have more distinct levels of proficiency. There is clearly a difference between the IEI learners and the 401 level learners, but there is some overlap between the L2 proficiency of the undergraduate and the graduate students. We might have more confidence in the reliability of the corpus if we also had yet another level of proficiency beyond the 401 students to see if there is any decrease in the persistent error categories like 2, 7 and 8. Also, some of the data are not terribly convincing because only one or two learners produced them. Notice, for example, five out of the seven infinitive-for-gerund errors recorded in the compositions of the ESL 114/115 learners were produced by one person. This argues the need for presenting this kind of data as a part of the total picture when reporting corpus based studies rather than just percentages alone. However, in one way, this corpus of Spanish-speaking learners of English is very suitable for L2 research – it is comprised of a variety of different topics and disciplines. The fact that it consists of written rather than spoken data is not a serious limitation, because we can assume that in writing the L2 learners have sufficient time to edit out any performance errors. This medium is therefore a valid representation of their IL grammars.

4. Conclusion

In this paper we have attempted to demonstrate some of the characteristics that L2 corpora must have if they are to be useful in researching aspects of the development of ILs over time. In addition to the criteria discussed in Granger (2002), such as diversity of topic, these characteristics must include multiple levels of proficiency and requisite sample sizes at each level. Furthermore, we have suggested that the use of corpus data to support arguments in favour of a particular hypothesis or as motivation for undertaking specific investigations must take into account the number of tokens produced by single individuals rather than just gross percentages of occurrence of particular syntactic phenomena. Our remarks should not be seen as implying that there are no researchers who utilise corpora to develop interesting theories of SLA. The valuable research of Meisel, Pienemann and Clahsen is a superb example of this. Because their research program focuses on determining stages, they do not feel that a strict determination of proficiency between subjects is absolutely necessary or relevant.

A recent study by Meisel (1997) utilised corpora to investigate cross-linguistic development of negation, but the description of the corpora and the subjects' proficiency are not extensively detailed, presumably because, once again, the focus was on determining overall patterns. However, in examining other questions, such as the extent to which L1 transfer errors are reflected in IL development, assessment of proficiency takes on a greater importance. As we have indicated, structured L2 corpora provide important database for exploring current issues in SLA, such as U-shaped development, stabilised IL grammars and the interaction of the L1 in the movement of ILs through various developmental stages. These data complement the use of SLA research methods such as experiments that employ grammaticality judgment and production tasks. They also facilitate cross-linguistic comparisons that allow us to expand our knowledge of whether certain error types are typologically determined or whether they are more widespread. This kind of information in turn enables researchers to determine with greater objectivity the nature and validity of SLA concepts such as transfer, overgeneralisation and stabilised grammars. Of course, the construction of large-scale L2 corpora that embody the above criteria will require cooperation between many researchers of different theoretical persuasions investigating IL development in typologically different languages. Nevertheless, such scientific cooperation should be doable, given the many corpus linguistics projects in various parts of the world.

References

- Anderson, R. (1991), 'Developmental sequences: the emergence of aspect marking in second language acquisition', in: T. Huebner and C. A. Ferguson (eds.) *Cross-currents in second language acquisition and linguistic theories*. Amsterdam: John Benjamins. 305-324.
- Anderson, R. and Y. Shirai (1994), 'Discourse motivation for some cognitive acquisition principles', *Studies in second language acquisition*, 16: 133-156.
- Anderson, R. and Y. Shirai (1996), 'The primacy of aspect in first and second language acquisition: the pidgin-creole connection', in: W. C. Ritchie and T. K. Bhatia (eds.) *Handbook of second language acquisition*. San Diego (CA): Academic Press. 527-570.
- Bardovi-Harlig, K. (1998), 'Narrative structure and lexical aspect: conspiring factors in second language acquisition of tense-aspect morphology', *Studies in second language acquisition*, 20: 471-508.
- Bardovi-Harlig, K. (1999), 'From morpheme studies to temporal semantics: tense-aspect research in SLA', *Studies in second language acquisition*, 21: 341-382.
- Bardovi-Harlig, K. and A. Bergström (1996), 'The acquisition of tense and aspect in SLA and FLL: a study of learner narratives in English (SL) and French (FL)', *Canadian modern language journal review*, 52: 308-330.

- Bardovi-Harlig, K. and D. Reynolds (1995), 'The role of lexical aspect in the acquisition of tense and aspect', *TESOL quarterly*, 29: 107-131.
- Biber, D., S. Johansson, J. Leech, S. Conrad and E. Finegan (1999), *Longman grammar of spoken and written English*. Essex: Pearson Education Ltd.
- Birdsong, D. and M. Molis (2001), 'On the evidence for maturational constraints in second-language acquisition', *Journal of memory and language*, 44: 235-249.
- Collins, L. (2002), 'The roles of L1 influence and lexical aspect in the acquisition of temporal morphology', *Language learning*, 52 (1): 43-94.
- Cowan, R., H.-E. Choi and D.-H. Kim (2003), 'Four questions for error diagnosis and correction in CALL', *CALICO*, 20: 451-463.
- Doughty, C. J. (2003), 'Instructed SLA: constraints, compensation, and enhancement', in: C. J. Doughty and M. L. Long (eds.) *The handbook of second language acquisition*. Oxford: Blackwell. 256-310.
- Ervin, S. (1964), 'Imitation and structural change in children's language', in: E. H. Lennenberg (ed.) *New directions in the study of child language*. Cambridge (MA): MIT Press. 163-189.
- Granger, S. (2002), 'A bird's-eye view of learner corpus research', in: S. Granger, J. Hung and S. Petch-Tyson (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: John Benjamins. 1-33.
- Han, Z.-H. (2000), 'Persistence of the implicit influence of the NL: the case of the pseudo passive', *Applied linguistics*, 21: 55-82.
- Kellerman, E. (1985), 'If at first you do succeed...', in: S. Gass and C. Madden (eds.) *Input in second language acquisition*. Rowley (MA): Newbury House. 345-353.
- Lardiere, D. (1998a), 'Case and tense in the "fossilised" steady state', *Second language research*, 14: 1-26.
- Lardiere, D. (1998b), 'Dissociating syntax from morphology in a divergent L2 end-state grammar', *Second language research*, 14: 359-375.
- Lardiere, D. (2000a), 'Mapping features to forms in second language acquisition', in: J. Archibald (ed.) *Second language acquisition and theory*. Malden: Blackwell. 102-179.
- Lardiere, D. (2000b), 'On optionality and grammaticality in L2 language', Paper presented at the 4th Meeting on generative approaches to second language acquisition (GASLA), MIT, Boston, USA, 1 april 2000.
- Long, M. (2003), 'Stabilisation and fossilisation in interlanguage development', in: C. J. Doughty and M. L. Long (eds.) *The handbook of second language acquisition*. Oxford: Blackwell. 487-535.
- MacLaughlin, B. and R. Hereda (1996), 'Information-processing approaches to research on second language acquisition and use', in: W. C. Ritchie and T. K. Bhatia (eds.) *Handbook of second language acquisition*. San Diego: Academic Press. 213-238.

- Meisel, J. (1997), 'The acquisition of the syntax of negation in French and German: contrasting first and second language development', *Second language research*, 13: 227-276.
- Norris, J. and L. Ortega (2003), 'Defining and measuring SLA', in: C. J. Doughty and M. L. Long (eds.) *The handbook of second language acquisition*. Oxford: Blackwell. 717-761.
- Pienemann, M., M. Johnston and G. Brindley (1988), 'Constructing an acquisition-based procedure for second language assessment', *Studies in Second Language Acquisition*, 10 (2): 217-243.
- Pinker, S. (1984), *Language learnability and language development*. Harvard: Harvard University Press.
- Roeper, T. (1981), 'In pursuit of a deductive model of language acquisition', in: C. L. Baker and J. J. McCarthy (eds.) *The logical problem of language acquisition*. Cambridge (MA): MIT Press. 129-164.
- Segalowitz, N. (2003), 'Automaticity and second languages', in: C. J. Doughty and M. L. Long (eds.) *The handbook of second language acquisition*. Oxford: Blackwell. 382-408.
- Schwartz, B. D. and R. A. Sprouse (1996), 'L2 cognitive states and the full transfer/full access model', *Second language research*, 12: 40-72.
- White, L. (2002), 'Morphological variability in endstate L2 grammars: the question of L1 influence', in: B. Skarabela, S. Fish, and A. H.-J. Do (eds.) *Proceedings of the 26th annual Boston University conference on language development*. Somerville (MA): Cascadilla Press. 758-768.
- Wolf-Quintero, K., S. Inagaki and H.-Y. Kim (1998), 'Second language development in writing: measures of fluency, accuracy, and complexity', Technical Report No. 17. Honolulu: Hawai'i Second Language Teaching and Curriculum Center.
- Zobl, H. (1989), 'Canonical typological structures and ergativity', in: S. Gass and J. Schachter (eds.) *Linguistic perspectives on second language acquisition*. Cambridge: Cambridge University Press. 203-221.

This page intentionally left blank

The path from learner corpus analysis to language pedagogy: some neglected issues

Nadja Nesselhauf

University of Heidelberg

Abstract

The analysis of learner corpora is often expected to contribute to language pedagogy, but the question of how exactly results from learner corpus studies can be turned into suggestions for language teaching has received practically no attention so far. In this paper, the necessity of discussing this question is pointed out and some directions the discussion could take are outlined. While the typical result of a learner corpus study is the identification of a number of features that are particularly difficult for a certain learner group, it is argued here that suggestions for language teaching should not be based exclusively on the criterion of difficulty and that, moreover, the criterion of difficulty itself needs refining in order to be truly relevant. The discussion of these theoretical issues is based on results from a learner corpus study investigating the use of collocations by advanced learners.

1. Introduction

The analysis of learner corpora often has pedagogical aims in that the results from learner corpus studies are, either explicitly or implicitly, expected to contribute to the improvement of (foreign) language teaching. However, the question of how exactly such results can be applied in the field of language teaching has hardly received any attention so far. This paper is an attempt to demonstrate that a discussion of this question is essential, as the path from learner corpus analysis to language pedagogy is not as direct as sometimes appears to be assumed. An attempt will also be made to indicate some directions that such a discussion could take. The argument will be based on some results of a learner corpus study investigating collocations in the English of advanced German-speaking learners of English (Nesselhauf 2005).¹ In Section 2, the aims and methods of the study will be outlined and some results presented. In Section 3, the question of how features of language can be selected for teaching on the basis of results from learner corpus studies will first be addressed in general terms and a possible path tentatively be outlined (Section 3.1). Then, the usefulness of the suggested path is tested with results from the study (Section 3.2).

2. A study based on a learner corpus

2.1 Aims, scope and methodology

Collocations in learner language appear to be a worthwhile object of study, as collocations are, on the one hand, pervasive in language, and, on the other, difficult even for advanced learners of English. In the literature on the topic, statements such as the following abound: “Any analysis of students’ speech or writing shows a lack of [...] collocational competence” (Hill 2000: 49). However, while the pervasiveness of collocations in language has been demonstrated in numerous studies, many of them corpus-based, evidence for the difficulty of collocations for learners is to a large degree anecdotal so far. The study reported here intends to shed light on the nature and extent of the difficulties of learners with collocations on the basis of empirical analyses. The learner corpus chosen for this purpose is the German subcorpus of *ICLE*, which contains argumentative and descriptive essays composed by advanced German-speaking learners of English. I used a slightly reduced version of this corpus, which I termed *GeCLE* (for *German ICLE*), and which contains around 300 essays or around 150,000 words (for details on *ICLE* cf. Granger 1996; for details on *GeCLE* cf. Nesselhauf 2005).

Collocations are defined here in a phraseological rather than in a frequency-based sense. This means that collocations are not defined as frequently co-occurring words, but as combinations which are not entirely predictable semantically but are arbitrary to some degree. According to this approach, on the one hand, a combination such as *drink tea* would not be considered a collocation, although the two words are likely to frequently co-occur. On the other hand, combinations such as *make a decision*, *take something into consideration*, *get in touch with somebody* would count as collocations, as part of the reason why these particular verbs and nouns are combined is arbitrary convention (why not, for example, **draw something into consideration* as in German?). The study is restricted to verb-noun combinations like the ones just cited.

Two types of verb-noun collocations are distinguished. This distinction is made according to the degree of restriction of a combination, and is therefore not to be considered a rigid distinction. One type of collocations will be referred to as RC2 (for Restricted Collocation Type 2). It consists of a verb that, in the sense in which it occurs in the collocation, can be combined with a fairly large number of nouns, but that does not combine with some nouns that seem equally plausible from a semantic point of view. For example, for the verb *reach* in the sense of ‘succeeding in achieving something’, the combinations *reach a decision/a conclusion/a compromise/an agreement/a goal* and a number of others are perfectly natural, whereas the combination *reach an aim* is not. The second type contains a verb that, in the sense in which it occurs in the collocation, can only be combined with a very small number of nouns; this type will be referred to as RC1 (Restricted Collocation Type 1). An example is *run a risk*, where the verb in the given sense is sometimes also found in combination with *danger*, but other

plausible combinations (such as *run a chance* or *run the peril of*) do not seem to be possible.

After manual extraction of all verb-noun collocations from *GeCLE*, their degree of acceptability was determined. This was done with the help of three types of sources (dictionaries, the *British National Corpus*, and native speaker judgements) and resulted in a five-step scale of acceptability from perfectly acceptable to completely unacceptable: +, (+), ?, (*), and *. For ease of reference, collocations judged '+' and '(+)' are referred to as 'acceptable' in what follows, those judged '*' and '(*)' as 'unacceptable', those judged '?' as questionable, and the unacceptable and questionable ones taken together as 'inappropriate' or 'deviant'.

2.2 Some results

Altogether, slightly more than 2000 verb-noun collocations were found in *GeCLE*. As Table 1 shows, about a quarter of these were judged to be unacceptable, and about a third were judged inappropriate, which supports the assumption that advanced learners have difficulties in this area of language.

Table 1: Acceptability of the verb-noun collocations in *GeCLE*.

+	(+)	?	(*)	*	total
1191	143	241	209	298	2082

The different types of deviations occurring in the collocations found in the corpus are listed in Table 2, together with the frequency of each type and an example each (with corrections in brackets).²

Table 2: Types of deviations in verb-noun collocations in *GeCLE*.

type of deviation	example	total
verb	make an experience (have)	389
noun	make a cut (distinction)	174
determiner	pass one's judgement (pass judgement)	35
noun complementation	have the possibility to+inf. (of - ing)	62
preposition in prepositional phrase	get into contact with (in)	8
support verb construction for verb	have a breakdown (break down)	39
whole collocation inappropriate	[To] come to an end ([To] conclude)	110
structure	set sb. an example (set an example for sb.)	19
total		836

Table 2 shows that the most frequent deviations by far occur in the verb, followed by deviations in the noun and instances where the whole collocation is inappropriate (for a detailed discussion of the different types of deviations see Nesselhauf 2005). Some of the examples given in Table 2 illustrate a more general finding of the study, namely that many of the collocations produced by the learners were not deviant as such, but merely used inappropriately in the context. *Make a cut*, *have a breakdown*, or *come to an end*, for example, are perfectly natural collocations, but *to come to an end* is used to conclude an essay, *have a breakdown* is used with reference to a computer, and *make a cut* occurs in the sentence *people have to understand that they have to make a sharp cut between science fiction [...] and reality*. A further result was that collocations that have both a similar form and meaning in L2 appear to be particularly liable to confusion. Examples are pairs or groups of expressions such as *get in contact – come into contact* and *make a difference – make a division – make a distinction*.

In the study, an attempt was also made to isolate factors that correlate with the difficulty of certain (types of) collocations. Factors investigated include the syntactic pattern of a collocation and whether it belongs to the category of support verb constructions. Two factors in particular yielded interesting results, the degree of restriction of a collocation and the congruence of a collocation in L1 and L2.

To investigate the correlation of degree of restriction and difficulty, the acceptability of the collocations of the types RC1 and RC2 were compared. The results are displayed in Tables 3 and 4, which show that while around 20% of the RC1 collocations in the corpus are deviant, this is true for around 28% of the RC2 collocations. The RC2 type (i.e. collocations with verbs that combine with a fairly large number of nouns but are nevertheless arbitrarily restricted in their combinability) thus appears to be more difficult for learners than the RC1 type.³

Table 3: Acceptability of RC1 collocations.

<i>acceptability</i>	+/(+)	?/(*)/*	<i>total</i>
number	259	63	322
percentage	80.4%	19.6%	100%

Table 4: Acceptability of RC2 collocations.

<i>acceptability</i>	+/(+)	?/(*)/*	<i>total</i>
number	1075	421	1496
percentage	71.8%	28.2%	100%

The second factor to be discussed here, congruence, refers to the possibility of translating a given collocation word-for-word from L2 into L1. If such word-for-word translation is possible (as for example in *have the right to+inf – das Recht haben zu+inf*), the collocation is referred to as ‘congruent’; if it is not possible (as for example in *make a decision – eine Entscheidung treffen*, but **eine*

Entscheidung machen), the collocation is referred to as ‘non-congruent’. The results of this analysis are presented in Tables 5 and 6.

Table 5: Acceptability of congruent collocations.

<i>acceptability</i>	+/(+)	?/(*)/*	<i>total</i>
number	933	190	1123
percentage	83.1%	16.9%	100%

Table 6: Acceptability of non-congruent collocations.

<i>acceptability</i>	+/(+)	?/(*)/*	<i>total</i>
number	401	292	693
percentage	57.9%	42.1%	100%

The tables show that around 17% of the congruent collocations are deviant, whereas a much greater percentage, namely 42%, of the non-congruent collocations is deviant. This means that if the concept a learner wants to express can be expressed by an English collocation which is a word-for-word translation of a German collocation, the likelihood that an acceptable collocation is produced is much greater than if the English and German collocations are non-congruent.⁴

3. The path to language pedagogy

3.1 Selecting features of language for language teaching

Results from learner corpus studies such as those I have just presented often have important implications for the question of how a language should be taught. For example, the finding that the collocations that the learners produced are frequently not unacceptable *per se* but rather are existing English collocations used inappropriately in the context should probably lead to the conclusion that collocations should not merely be taught in isolation but also in context and that attention should be paid to their usage. What I would like to focus on here, however, is the question of what should be taught, i.e. on the selection of language features for language teaching. More generally, I would like to discuss how results of the type ‘feature A is more difficult than feature B’, which are the typical results of a learner corpus study, can be translated into suggestions for language teaching.

At the moment, there appears to be a tendency for learner corpus studies to conclude by stating which features were found to be particularly difficult for a certain learner group (i.e. to be deviant particularly often) and, therefore, by asserting that the emphasis on this feature in teaching should be increased. In other words, the equation applied is as follows:

the more difficult = the more emphasis required in teaching

A similar tendency has been noted to exist for studies based on native speaker corpora which aim to contribute to the improvement of language teaching. The equation usually applied in those cases appears to be:

the more frequent = the more emphasis required in teaching

This latter equation, which relates native speaker corpus analysis to language teaching, has frequently been criticised (e.g. Widdowson 2000). The first equation, referring to the relation of learner corpus analysis to language teaching, has not received much criticism, despite the fact that, like the second one, it also bases the selection of features for language teaching on one criterion exclusively. Therefore, it also has to be considered as too simplistic.

If the literature on language teaching is consulted on the question of the selection of vocabulary, the range of criteria found is usually much broader. These theoretical suggestions usually resemble the following list of criteria provided by Nation (1990, based on Richards 1970):

- frequency
- range
- language needs
- availability and familiarity
- coverage
- regularity
- ease of learning or learning burden

However, such lists often appear to have been drawn up with the beginner or intermediate learner in mind and need to be revised for the teaching of advanced learners, who are the focus of attention here (and tend to be the focus of learner corpus studies in general). A feature such as ‘ease of learning’, for example, i.e. whether a certain feature is easy or difficult to acquire, should probably not play a major role when selecting features of a language for teaching advanced learners. Similarly, the features of coverage and regularity, which refer to the frequency with which one expression can replace others and to whether a feature is more or less regular, do not seem of central importance for this group of learners. On the other hand, the criterion of difficulty (i.e. the degree to which an expression is likely to be deviant), which does not figure at all on the above list, certainly is a major criterion. The same applies to frequency, although this criterion needs some refining. Frequency probably ought to be combined with what Nation refers to as ‘range’ and by which he means frequency over many text types: it would appear sensible to select items of language for (advanced) teaching if they are frequent in the text types the learner needs to be familiar with.

A further criterion relevant for the advanced learner, which appears to be missing in Nation’s list, is the criterion of disruption that arises if a certain item is used inappropriately. Disruption refers to the degree to which an expression, if deviant, draws the listener’s or reader’s attention away from the message or

makes them misunderstand or fail to understand the message. As a preliminary suggestion, I would therefore propose that the three major criteria for the selection of vocabulary (and other language items) for advanced courses should be:

- frequency
- difficulty
- degree of disruption

The next step would be to decide how exactly each of these three criteria can be measured and how they should be weighted. For the criteria of frequency and degree of disruption, these questions will be left open here, in full realisation of the fact that the latter is particularly problematic (as it is almost impossible to measure disruption objectively, and as questions such as disruption for whom and under what circumstances would have to be addressed). For the criterion of difficulty, some issues shall be raised here, however, as this concept figures so prominently in learner corpus analysis. Indeed, it needs to be pointed out that difficulty is usually not to be understood in any psycholinguistic sense in learner corpus analyses but mostly refers to the degree to which a combination is susceptible to deviation. Although there appears to be wide agreement on this aspect of the term's use, the concept of difficulty often remains hazy in other respects in learner corpus analyses.

Two types of difficulty can theoretically be studied in an analysis of learner data: absolute and relative difficulty. In the former case, the absolute number of deviant items in a given amount of text is counted. In the latter case, the number of deviant items of a certain kind is related to the number of times this item was correctly produced. This distinction is only very rarely made or pointed out in learner corpus studies, and it is even rarer that both kinds of difficulty are examined. Nevertheless, in order to determine the overall degree of difficulty of an item, both its relative and absolute difficulty should be taken into account. Features have clearly not been adequately acquired no matter whether the feature is deviant, say, 10 times in a given amount of text and produced correctly 30 times, or whether it is deviant only 3 times but never produced correctly. The question of how these two types of difficulties should be weighted will also require some attention in the future.

A further question that needs to be addressed with respect to the criterion of difficulty is which elements should be considered difficult if a deviation occurs: if a learner produces feature A instead of feature B, does that mean he/she is experiencing difficulties with feature A or with feature B? Again, account probably needs to be taken of both. The learner has obviously not acquired feature B in a way that leads to its production when it is called for and at the same time has not acquired feature A in a way that leads to its appropriate use. The question of how precisely these two aspects can be integrated in the path from analysis to suggestions for teaching also clearly needs to be addressed by future research.

3.2 Applying the theory to the results of the study

When trying to apply these theoretical considerations to some of the results of the study previously outlined, it becomes immediately clear that they can be applied best to results that can be quantified relatively precisely. In the case of the result that collocations with both a similar form and meaning in L2 are particularly difficult (which is a rather impressionistic result that can hardly be quantified), for example, neither absolute nor quantitative difficulty can be determined exactly. What can be applied nevertheless, albeit only roughly, is the two criteria that have been set up in addition to difficulty. For the pairs and groups of collocations identified as particularly liable to confusion (such as *make a difference – make a distinction – make a division*, cf. Section 2.2), frequency in native speaker language would need to be investigated individually. In general, however, many of the pairs and groups identified appear to be frequent in English. With respect to the criterion of disruption, it can probably be assumed that misunderstandings can arise easily if two similar collocations are confused, as the form of what is produced is correct and might therefore not immediately be noticed as deviant by the listener. If this assumption is borne out by further investigation, this particular result could then with some justification lead to the suggestion that those groups collocations that are both similar in form and meaning and frequent in English deserve particular attention in teaching.

In those cases where results can be quantified relatively precisely, in order for the path suggested above to be followed, two requirements would ideally have to be met: learner corpus analysis would have to proceed with increased precision and should be supplemented both by native speaker corpus analysis of the same phenomenon and by investigations of disruption. For the sake of illustration, the latter two points will be covered by rough estimates in the following discussion.

Let us first consider the result that non-congruent collocations are more difficult than congruent collocations. The results presented above (in Tables 5 and 6) refer to the relative difficulty of these two types of collocations (42% vs. 17%). In absolute terms, non-congruent collocations are also more difficult than congruent ones, but the difference in the absolute difficulty of these two types is smaller than the difference in their relative difficulty: altogether, 292 deviant non-congruent collocations and 190 deviant congruent collocations were identified. In addition, the results refer to the collocations that the learners apparently intended, i.e. that were required in the context. For example, if a learner produced *make an experience* instead of *have an experience*, this is considered a deviation in a non-congruent collocation, as the apparent target *have an experience* and the equivalent German collocation *eine Erfahrung machen* are non-congruent. In such cases, where what was produced is actually non-existent in English, this seems the only sensible approach, as the question of whether the collocation actually produced is also difficult does not arise.

Another example is the production of *make a difference* instead of *make a distinction*, which was classified as a deviant non-congruent collocation (as *make a distinction* and *eine Unterscheidung treffen* are non-congruent). With equal

justification, this deviation could, however, also be considered a problem with a congruent collocation (*make a difference – einen Unterschied machen* being congruent). In fact, it constitutes a problem with both a congruent and a non-congruent collocation, although this is not reflected in the results presented above. The focus on the apparently intended collocations – problematic as this concept may be (as there is no way of knowing what a learner actually intended) – was chosen because it was deemed of particular importance to find out whether the fact that the learner's concept can or cannot be expressed in the L2 like in the L1 has an influence on collocation production. As the above example shows, however, such a one-sided approach to difficulty is insufficient and needs to be supplemented by taking into account the items actually produced.

As to the additional criteria of disruption and frequency, it can be assumed that, since congruence is an inter-language phenomenon, there is no correlation between disruption and congruence. As to frequency, it seems likely that for the language pair under investigation (English and German), congruent collocations are more frequent. Taken together with the results on relative and absolute difficulty (and provided these assumptions are accurate), this would mean that the suggestion of putting a somewhat greater emphasis on non-congruent collocation than on congruent ones in teaching is justified.

If we now look at the result that RC2 collocations are more difficult than RC1 collocations, the first point to be noted is that (as in the case of congruence) only relative difficulty and only the apparently intended collocations were considered. Again, this would need to be supplemented by also considering the collocations actually produced and the absolute difficulty of the two types. In contrast to the numbers obtained for the factor of congruence, there is a considerable difference in difficulty between RC2 and RC1 collocations in absolute terms as well: 421 deviant RC2 collocations can be identified, as opposed to 63 deviant RC1 collocations. The factors of difficulty of actually produced collocations and of disruption will be left unquantified again, although it can perhaps be assumed that disruption does not generally correlate with the two groups of collocations. As to frequency, there can hardly be any doubt that RC2 collocations are more frequent in English than RC1 collocations (as the verbs combine with many more nouns). Together with the fact that both relative and absolute difficulty are considerably greater for RC2 collocations than for RC1 collocations, the suggestion that RC2 collocations should receive considerably more emphasis in teaching than RC1 collocations therefore seems justified.

If the two factors of congruence and degree of restriction are compared, this means, then, that the degree of restriction should be an even more important factor than congruence in the selection of collocations for teaching (unless, of course, one of the assumptions made here is wrong). First, the difference in absolute difficulty is far greater for RC2 and RC1 collocations than for non-congruent and congruent collocations. Secondly, the criterion of frequency points in the same direction as the criterion of difficulty in the case of RC1 and RC2 collocations but not in the case of congruent and non-congruent ones. Had we

directly relied on the results of the learner corpus analysis as presented in the previous section when making suggestions for teaching, the conclusion would have been the opposite, namely that congruence should be a more important factor in the selection of collocations for teaching than degree of restriction.

To conclude, the discussion presented here has demonstrated that by taking into account more than merely one criterion for the selection of language features in teaching and by additionally refining the criteria, much better founded suggestions can be made than if only one (aspect of one) criterion is taken into account. I also hope to have shown that for learner corpus analysis to become truly relevant for the field of language teaching, further work on the development of a sensible path from learner corpus analysis to language pedagogy is necessary.

Notes

- 1 In Nesselhauf 2005, both the methodology and the results of the study are reported in much greater detail than is possible here.
- 2 The overall number of individual deviations (836) is higher than the number of deviant collocations, as in several collocations more than one type of deviation occurred.
- 3 The reason that the totals of the two tables do not add up to the overall number of collocations found in the corpus is that only those cases were considered where a collocation was both produced and apparently intended (i.e. required in the context). Cases where an RC1 or RC2 collocation had been produced but an expression which was not a collocation, as for example a simple verb, was required in the context, were excluded (cf. also Section 3).
- 4 As in the previous analysis, only those cases where a collocation was both produced and apparently intended were considered.

References

- Granger, S. (1996), 'Learner English around the world', in: S. Greenbaum (ed.) *Comparing English worldwide: the International Corpus of English*. Oxford: Clarendon. 13-24.
- Hill, J. (2000), 'Revising priorities: from grammatical failure to collocational success', in: M. Lewis (ed.) *Teaching collocation. Further developments in the lexical approach*. Hove: LTP. 47-69.
- Nation, I. S. P. (1990), *Teaching and learning vocabulary*. Boston: Heinle & Heinle.
- Nesselhauf, N. (2005), *Collocations in a learner corpus*. Amsterdam: John Benjamins.

- Richards, J. C. (1970), 'A psycholinguistic measure of vocabulary selection', *International review of applied linguistics*, 8: 87-102.
- Widdowson, H. (2000), 'On the limitations of linguistics applied', *Applied linguistics*, 21: 3-25.

This page intentionally left blank

Exploiting the Corpus of East-African English

Josef Schmied

Chemnitz University of Technology

Abstract

This contribution summarises the current status and future plans for the Corpus of East African English (ICE-EA), which is part of the International Corpus of English¹ (cf. Greenbaum 1996, Schmied 1996 and forthcoming). It concentrates on methodological problems that are related to the exploitation tools on several discovery levels, starting from the individual corpus text through the entire ICE-EA up to the World-Wide Web as a corpus (Schmied 2005). It shows that the investigations using ICE corpora for comparative theoretical and practical analyses of second-language varieties of English has hardly started, but also that interesting new research opportunities are available.

1. Linguistic issues for ESL corpora

Corpora of English as a second language (ESL) are – in contrast to EFL or learner corpora – characterised by their own special features on every level of linguistic analysis, which are not interference phenomena but due to the special learning process in formal settings (e.g. the importance of intralanguage features like generalisation or regularisation, and the strong influence of formal, written usage). This ranges from the lexicon (Schmied 2005) to the syntactic, semantic and pragmatic levels.

Syntactically, for instance, complexity has been a major issue for the analysis of ESL corpora. Usually ESL varieties are supposed to be syntactically less complex, since they are related to learner varieties of English which have not developed the most sophisticated forms of linguistic codification. However, since ESL corpora are also associated with special prestige in their sociolinguistic context, some speakers and thus some texts are also characterised by more complex structures than ENL texts. In an untagged corpus this can be analysed on the basis of conjuncts or prepositions, which mark openly clausal and noun-phrase complexity. In a tagged corpus many more structures can be analysed on the noun- as well as the verb-phrase level, particularly modification clusters and clause types.

Semantically, ESL varieties are usually characterised by specific selection restrictions and stereotypes on the one hand and collocational flexibility on the other.

Pragmatically, certain style features of African and Asian cultures have been reported as part of ESL varieties, particularly conversation styles including greetings and culture-specific address forms.

In the long run, the *ICE* corpora will enable us to compare native- and non-native (ENL and ESL) corpora world-wide. The corpora and tools available now show how fruitful such a comparative perspective can be (Schneider 2004 or Sand 2004). Recently, a very rough 'research guide' has been sketched out (Fallon 2004), which is based on secondary literature. What is missing now is a sketch of variationist research issues. The following section will at least show the general way to achieve this. Since the *ICE* project has always emphasised the wide and, as far as possible, standardised stratification of text-types, it is the ideal testing ground for such variationist issues, since it has been mentioned in almost all detailed studies that variation across text-types is more pronounced than variation across varieties.

2. Research hypotheses for ESL varieties and some empirical evidence

The possible exploitation of *ICE-EA* (as well as other *ICE* corpora) can be based on the following hypotheses:

1. *ICE-EA* is too small for many collocational/distribution analyses, but it shows the limitations of the *ICE* approach and serves as a starting point for detailed WWW searches;
2. the lexical complexity of *ICE-EA*/ESL varieties is restricted (e.g. type/token relationship), but also enriched by East African lexemes;
3. the syntactic complexity of *ICE-EA*/ESL is reduced, but increased in some (prestige-related) text-types like broadcast discussions;
4. the idiomaticity and the collocates are often conventionalised (stereotyped), but the forms are more flexible in *ICE-EA* and possibly ESL varieties in general.

These hypotheses cannot be proven on the basis of the data available, but only illustrated. The corpus exploitation tools necessary are *either* part of the *ICE* corpus, general corpus tools like *WordSmith* or new tools based on the concept of the World Wide Web (WWW) as a dictionary² (cf. Schmied 2005). Thus hypothesis 1 can be illustrated on the basis of *ICE-EA*. The fairly well-known and common East Africanisms *matatu* (often translated as 'collective taxi' in East Africa or 'taxi bus' in other countries) can be shown as a clear East Africanism occurring in a relatively wide variety of texts (44 times in 20 texts) in spoken and written Kenyan English – but not in Tanzanian English.

Search Results										Help					
Statistics															
Exact phrase hits for:		matatu					44								
Relative Statistics (normalized to total number of words)					Absolute Statistics										
Domain	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	percent normalized	percent absolute	hits	result text words	total words
Kenya written	██████████										44.33 %	43.18 %	19	19,907	377,111
Tanzania written											0.00 %	0.00 %	0	0	398,035
Kenya spoken	██████████										55.67 %	56.82 %	25	16,970	395,081
Tanzania spoken											0.00 %	0.00 %	0	0	217,710
												44	36,877	1,387,937	
Texts that contained a hit (result texts):												20			
Number of all examined texts (total texts):												953			

Figure 1: Search results for *matatu* in ICE-EA.

Similarly, the keyword *ugali* in East Africa can be shown to occur in several contexts and cotexts, which give at least a clue to its meaning. The keyword in context analysis (Fig. 2) shows us that *ugali* can be seen parallel to *rice* as a main base of a traditional African dish; it also shows that it has to be *prepared* or *made*, with *water* and that it is fairly well integrated in East African English, since it is hardly translated or explained in an apposition etc.

0.0165575112735955 hits per 1000 words

br-talkK.txt:Kenya / spoken / monologue / scripted / broadcast / talks

Line 124 you do like a gourd of water or hot *ugali* <

column-T.txt:Tanzania / written / printed / persuasive / personal / columns

Line 16 especially when it comes to hiding those chunks of *ugali* and ubwabwa dishes.

Line 135 provides lunch and supper to her customers by preparing *ugali*, rice,

convers1-K.txt:Kenya / spoken / dialogue / private / conversations

Line 9 ea/>Si my *ugali*

Line 12 C> We want to talk about my *ugali* and why men cook

Line 440 ea/>akasema hapana now this babe thinks this *ugali* will be enough Then he saw another drum there this ..just poured the thing So the babe was making *ugali* uh

Line 2174 O/> fits with everything else rice *ugali*

Line 2215 fact as I talk I'm some pieces of *ugali* <O/> So I told the guy So when the maji

Line 2217 The *ugali* was a kilo in fact more 'cause of the water ...

Figure 2: Search results for *ugali* in ICE-EA.

The term *ugali* occurs 21 times in 14 *ICE-EA* texts, but that figure appears too low for further analyses (e.g. to ask why it occurs most often in spoken Kenyan English but least often in spoken Tanzanian). The fact that *ugali* is well integrated in East African English as a whole is confirmed by a WPC analysis, which includes Uganda (Table 1).

Table 1: WPC search results for phrase *ugali*.

	<i>domain</i>	<i>hits</i>	<i>total pages</i>	<i>intra-domain %</i>	' <i>relevance</i> ' %
1	.ke	10	442	2.262	43.774
2	.tz	5	325	1.538	29.766
3	.ug	5	367	1.362	26.36
4	.za	33	755,000	0.004	0.085
5	.uk	134	15,800,000	0.001	0.016
	<i>total</i>	187	16,556,134	0.001	100

Although in absolute the term *ugali* occurs even more frequently in South Africa and the UK, it is clearly identified as an East Africanism by WPC.

If we compare this table with the results of a search in *WebCorp* we see that on an international level *ugali* is often explained by adding *corn meal* in brackets or a similar apposition. However, *WebCorp* also reveals that none of the WWW pages are clearly marked as Kenyan or Tanzanian by the appropriate TLD (Top Level Domain, i.e. ke, or tz). It is also interesting that the collocation 'Cameroonian *ugali*' occurs, which seems to contradict our hypothesis that *ugali* is a typical East Africanism; however, it only shows the limitation of a WWW research, since the term is not used at all in Cameroon itself. The statistical search results (Table 1 above) for *ugali* using the TLDs ke, tz, ug, za and uk shows that although the individual hits may be rather small, they are clearly very common in the three East African countries and only used very specifically in the much larger domains for South Africa and the United Kingdom, which is made very clear through the intra-domain percentage, which sets the search lexeme in relation to the size of the WWW site (based on occurrence of the English *the*).

More detailed analyses of syntactic complexity in East African English have to be carried out on the basis of a tagged corpus (Fig. 3):

S1B021K_FO <\$A>_NULL I_PPIS1 am_VBM Goro_NP1 Kamau_NP1 Welcome_VV0 Well_RR to_TO begin_VVI with_IW I_ZZ1 would_VM like_VVI to_TO ask_VVI Dr_NNB Gikenye_NP1 to_TO tell_VVI us_PPIO2 something_PN1 in_RR21 brief_RR22 what_DDQ </>_NULL ma_NN1 medicine_NN1 is_VBZ all_DB about_RP <\$B>_NULL Thank_VV0 you_PPY very_RG much_DA1 Medicine_NN1 is_VBZ an_AT1 aspect_NN1 in_II life_NN1 that_CST deals_VVZ with_IW those_DD2 areas_NN2 that_CST are_VBR concerned_JJ with_IW the_AT health_NN1 delivery_NN1 to_II the_AT community_NN1 Now_RT this_DD1 service_NN1 is_VBZ provided_VVN by_II various_JJ types_NN2 of_IO persons_NN2 who_PNQS were_VBDR trained_VVN in_II their_APPGE own_DA right_NN1 In_II this_DD1 instance_NN1 I_PPIS1 have_VH0 in_II mind_NN1 doctors_NN2 nurses_VVZ uh_UH pathologists_NN2 and_CC many_DA2 more_DAR others_NN2 <\$C>_NULL Okay_RR uh_UH to_II talking_VVG about_II medicine_NN1 well_RR you_PPY are_VBR you_PPY 've_VH0 defined_VVN medicine_NN1 but_CCB you_PPY focused_VVD more_RRR on_II the_AT on_II the_AT modern_JJ doctor_NN1 Well_RR could_VM you_PPY please_RR tell_VVI us_PPIO2 something_PN1 about_II uh_UH our_APPGE </>_NULL our_APPGE traditional_JJ the_AT traditional_JJ basis_NN1 of_IO modern_JJ medicine_NN1 <\$B>_NULL Well_RR I_PPIS1 really_RR am_VBM not_XX anywhere_RL near_II an_AT1 expert_NN1 in_II traditional_JJ medicine_NN1 and_CC I_PPIS1 think_VV0 I_PPIS1 would_VM not_XX be_VBI doing_VDG the_AT traditional_JJ </>_NULL med_VVD uh_UH medical_JJ man_NN1 medicine_NN1 man_NN1 a_AT1 service_NN1 but_CCB uh_UH suffice_VV0 it_PPH1 to_TO say_VVI that_CST there_EX is_VBZ a_AT1 difference_NN1 between_II the_AT delivery_NN1 of_IO health_NN1 care_NN1 by_II the_AT medical_JJ by_II the_AT modern_JJ medical_JJ man_NN1 and_CC by_II the_AT traditional_JJ man_NN1 traditional_JJ medicine_NN1 man_NN1 uh_UH now_RT the_AT traditional_JJ medicine_NN1 man_NN1 did_VDD his_APPGE training_NN1 by_II apprentice_NN1 He_PPHS1 learned_VVD from_II his_APPGE predecessor_NN1 and_RR31 so_RR32 forth_

Figure 3: Tagged *ICE-EA* spoken.

A quick superficial analysis of the type of noun phrase complexity expected shows that the co-occurrence of adjectives is more common than two adjectives in a row or even two adjectives combined by a conjunct in front of a general noun. In CLAWS tagging (which would make *ICE-EA* compatible to the *British National Corpus (BNC)*, the largest reference corpus available), this means *_JJ *_JJ *_NN1 >> *_JJ *_CC *_JJ *_NN1. For such detailed analysis, to confirm hypothesis 3, a much broader corpus of East African English would have to be collected. Nowadays this can be done on the basis of the world-wide web relatively quickly, and we have developed our own tool for that called *WebGrabber* (linked to an Ngrammer; cf. 4.1. below), which extracts n-long strings of continuous characters in a document. Thus syntactic chunks can be extracted and compared. For East Africa, first tests suggest that syntactic patterns are more frequently used than in comparable texts in native English varieties. This was also confirmed in a detailed study on prepositions in *ICE-EA* (Mwangi 2003), which demonstrates that less common complex prepositions (like *in view of*) are used even less frequently and the most common prepositions and the most frequent meanings (*in* with locative meanings) more frequently. These are special cases of the restricted lexical complexity of ESL varieties, which is balanced out by East Africanisms as culture-specific enrichments of English (hypothesis 2).

What can be done in a much more detailed lexicological analysis and subsequent lexicographical description of East Africanisms can be illustrated here for the term *matatu* (Fig. 4, which could be complemented by an appropriate image on the web easily):

matatu *pl* ~*s* *N* ‘collective taxi’, ‘taxi bus’ in EAfr., esp. Kenya

usu. licensed for fixed routes of public transport, but flexible, they leave when ‘full’;

infamous for reckless driving and overcrowding;

etym. <Sw. “three”, *orig.* 3 Shs fare;

collocates: *agent* driver, tout, operator, passenger; *locative* park, stand, stage, stop;

Prep. in, on board a ~; *Verb* enter, board

Figure 4: Lexical entry for *matatu*.

The general grammatical details (about the plural *-s* and the word class), the meaning (and its regional restriction to East Africa) and the etymology from the Swahili term for *three* (since originally the fare was three Shillings) can be included. However, to fully understand the importance of *matatus* for the Kenyan economy and thinking, more detailed explanations including associative meanings have to be added. For, today *matatu* can also be used as a noun premodifier in terms like *matatu mentality*, which always carries a clear negative connotation in Kenya.

In case grammar, the accompanying nouns for the term *matatu* can be described: the nouns *driver*, *tout*, *operator* and *passenger* are very common as agentives and *park*, *stand*, *stage* and *stop* as locatives; the prepositions *in* or *on board a matatu* are common collocates and finally the verbs *enter* or *board* are used, as in *enter/board a bus*.

This shows that a corpus-linguistic approach with its key-words-in-context is rewarding for a comprehensive lexical description of African lexemes and African usage. The problem of idiomatic flexibility (hypothesis 4) can be illustrated using the term *grass roots/grassroots/grass root* (Table 2).

Table 2: ICE-EA/ESL idioms are like *grass roots* less fixed/more flexible.

	Kenya			Tanzania			Σ
	Σ	written	spoken	Σ	written	spoken	
grass roots	4	3	1	16	10	6	20
grass roots	12	11	1	9	3	6	21
grass root		1				1	2
Σ	16			25			41

A comparison of the spoken and written parts of Kenya and Tanzania reveals some interesting differences: whereas *grassroots* written as one word seems to be more common in Tanzania, *grass roots* as two words is more often used in Kenyan, particularly written English. *Grassroot* is obviously only used as a nominal premodifier in both countries.

A broader comparison of East African lexemes and their (plural) variation can only be investigated on the basis of the WWW. Table 3 shows again the term *ugali* (*ugalis*) but also parallel *juakali* (*juakalis*) and *matatu* (*matatus*), *askari* (*askaris*). Here we can see again that, despite low absolute figures, the relative frequency of all East Africanisms is relatively high, even on the few web pages under the TLD .ke and .tz. Again, *WebCorp* gives us appropriate collocates, for instance, for the term *jua kali* we find the following key phrases:

<p>Key Phrases: <u>the jua kali</u> <u>a jua kali</u> <u>na jua kali</u> <u>informal jua kali</u> <u>for jua kali</u> <u>and jua kali</u> <u>wa jua kali</u> <u>jua kali sector</u> <u>jua kali firms</u> <u>jua kali economy</u> <u>jua kali industry</u></p>

Figure 5: Top external collocates of *jua kali*.

Similarly, the collocates of *jua kali* can be seen in Table 3:

Table 3: Collocates of *jua kali* in *WebCorp*.

<i>word</i>	<i>total</i>	<i>L4</i>	<i>L3</i>	<i>L2</i>	<i>L1</i>	<i>R1</i>	<i>R2</i>	<i>R3</i>	<i>R4</i>	<i>left total</i>	<i>right total</i>
The	44	5	4	1	27			3	4	37	7
Kali	28	3	8		2	3	2	6	4	13	15
And	21	2	2	1	4	1	8	1	2	9	12
In	19	2	1	4		1	10		1	7	12
Jua	17	6		3		3		4	1	9	8
Sector	16	1				12	2	1		1	15
To	14	6		1	2		5			9	5
Of	13	4		5			1		3	9	4
A	12			2	8		2			10	2
For	10	3	1		4			2		8	2
Na	9	1			5			1	2	6	3
The	9	1	2	3	3					9	0
informal	9			1	4	2	1		1	5	4
Ya	9	1	4	1	3					9	0
From	9	1	1	4				1	2	6	0
Jua	7	1		1		1	2		2	2	5
Firms	7		1	1		5				2	5
Artisans	6	1	1			2	1	1		2	4
By	6	1		4					1	5	1
As	6			3		1	2			3	3

The final example is the term *mitumba*, which is used extremely often nowadays in East Africa, but only started appearing more frequently after the special economic changes that made second-hand clothes a very well-known phenomenon in East Africa. The examples (Fig. 6) from *ICE-EA* show the usage, particularly that *mitumba* is often used with an explanation (e.g. in parentheses) in the early 1990s and that it is coded as an East Africanism <ea/> in the corpus, in

contrast to *parastatal*, which is also an East Africanism (though it seems hardly used in Kenyan spoken English), but it is not marked as East Africanism, as it is not perceived as such by East Africans.

<p>br-int-T.txt: S1BINT14T and for them first they used uh these used clothes <ea/><i>mitumba</i>.</p> <p>ppsocsc-T.txt: W2B019T a driver for a parastatal in the city says that <ea/><i>mitumba</i> (second hand clothes) are his saviour.</p> <p>pptech-T.txt: W2B036T for wearing apparel has led to the importation of cheap “<ea/><i>mitumba</i>” which in turn have practically squeezed out the industry...</p> <p>column-T.txt: W2E018T one, are trekking to work, have resorted to second-hand clothes (<ea/><i>mitumba</i>) etc. ...</p>
--

Figure 6: Search results for the East Africanism *mitumba* in *ICE-EA*.

Finally, the differences in spoken and written East African English can be demonstrated using the term *witchcraft* (Fig. 7).

Search Results										Help					
Statistics															
Exact phrase hits for:		witchcraft				95									
Relative Statistics (normalized to total number of words)										Absolute Statistics					
Domain	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	percent normalized	percent absolute	hits	result text words	total words
Kenya written											0.00 %	0.00 %	0	0	377,111
Tanzania written											65.98 %	75.79 %	72	5,530	398,035
Kenya spoken											5.54 %	6.32 %	6	6,219	395,081
Tanzania spoken											28.48 %	17.89 %	17	2,132	217,710
												95	13,881	1,387,937	
Texts that contained a hit (result texts):												7			
Number of all examined texts (total texts):												953			

Figure 7: Statistics search results for *witchcraft* in *ICE-EA*.

3. Cultural differences in ESL varieties like *ICE-EA*

These examples lead us to assume a cline of cultural differences in language varieties, in particular in ESL varieties, which can be summarised as follows:

1. The lexicon is clearly marked by cultural differentiation, particularly in some semantic fields like food and cloth items, which are usually clearly the domains of loans from first languages. Most of these items are common in Kenya and in Tanzania (*ugali* with a 12:9 ratio), whereas political terms like *ndugu* are only common in Tanzania (286:4).
2. Pragmatics is also an interesting level of culture-specific influence, particularly since many people are often unaware of it. Modality and specific address forms are clear examples for this phenomenon. Again detailed analyses are difficult without a tagged corpus. Thus *now* has been identified as a special African discourse maker (cf. Jeffrey and van Rooy 2004), but this has to be distinguished from other temporal usages and the statistics (Fig. 8) have to be subdivided.
3. Idiomaticity is an area where users are unaware of culture-specific influences, although metaphors and special collocations creep in very often.
4. Finally, grammar is, of course, only rarely an indicator of cultural variation, however, the use of modal verbs or of passives (for instance in gender-specific contexts) illustrates subtle differences here as well.




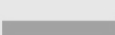
Statistics														
Exact phrase hits for:											now		2,564	
Domain	Relative Statistics										Absolute Statistics			
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	percent	hits	result text words	total words
Kenya written											15.52 %	398	252,542	377,111
Tanzania written											12.95 %	332	237,380	398,035
Kenya spoken											45.05 %	1,155	350,201	395,081
Tanzania spoken											26.48 %	679	200,366	217,710
											2,564	1,040,489	1,387,937	
Texts that contained a hit (result texts):											607			
Number of all examined texts (total texts):											953			

Figure 8: Statistics search results for *now* in *ICE-EA*.

Another much-quoted pragmatic example is provided by greeting formulae. This may explain the surprising idiosyncrasy in several social letters by different writers in *ICE-EA* like:

W1B-SK03 *How are* you *since* we departed on </29nt> March?

W1B-SK08 *How are* you *since* my last departure and sight at Nakuru

W1B-SK17 *How are* you *since* my last sight as well as departure.

W1B-SK46 *How are* you *since* last friday . I hope you are very fine and ...

W1B-SK47 Dear <name/>, So *how are* you *since* yesterday?

So far *ICE-EA* is available freely for everyone on the world-wide web as raw data. However, an annotated version would be much more useful as a special service, also to non-linguistic users, and could lead to a comparative ESL dictionary which would be interesting for ESL users in East Africa and world-wide. For East Africans, the special emphasis on collocates and metaphors would make them much more aware of their own usage, whereas for non-East Africans special paraphrases and even photos for specific lexical items would be an invaluable introduction to East African English and culture.

4. New ELL tools to expand the database

4.1 *WebGrabber*

WebGrabber allows the user to grab large amounts of data from the Internet. After selecting the required depth for the retrieval, the users can select which files they want to include into the retrieval process. *WebGrabber* offers all the files that are characterised as text files according to their extensions. This methodology allows a fast and effective retrieval of relatively large amounts of data which can be used as an Ad-hoc- or monitor corpus against which other data, for instance from specific *ICE* files can be evaluated as more or less representative for a specific text type or variety. For copy-right reasons the raw data cannot be stored permanently but they can be processed so as to be used for further linguistic analysis. One option is part-of-speech tagging, where the users can even choose from two different types of taggers: Qtag and the TNT-tagger; the latter is particularly useful because it is largely CLAWS compatible and this again relates our ad-hoc corpora to the *BNC* reference corpus.

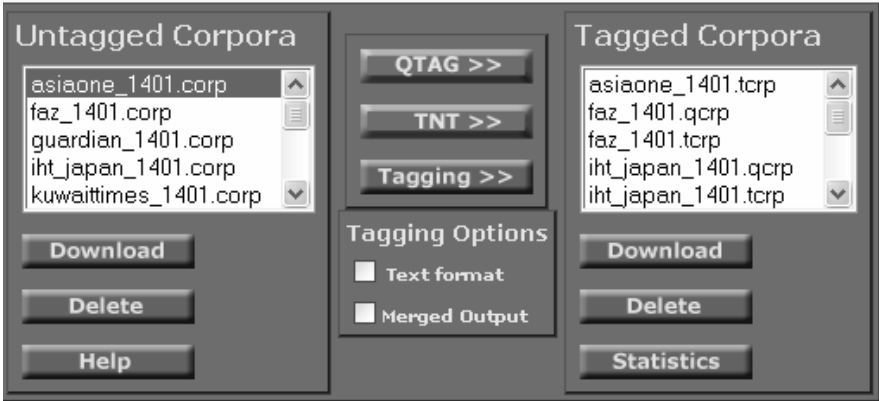


Figure 9: Tagging options in *WebGrabber*.

4.2 *ICE-Web*

ICE-Web as a monitor corpus allows the comparison of *ICE* corpora with a larger text collection which is more stratified and more adapted to the *ICE* criteria than the WWW as a corpus. *ICE-Web* is stratified according to regional and genre perspectives. The regional subcategorisation may be as diverse as in SE Asia, where English-using nations have as diverse a history as Burma and the Philippines or Indonesia and Vietnam. East Africa is relatively homogeneous, comprising the three states Kenya, Tanzania and Uganda, if North East Africa and Central Africa are kept separate. The genre classification comprises four big groups of text types, which correspond partly to *ICE* categories. Newspapers are customarily subdivided into the more objective reportage and the more subjective editorial, as well as the less professional letters to the editor. Similarly, radio broadcasts are divided into news and commentary. Business web pages can be divided at least into information and advertising, although more informational and more persuasive texts may, of course, overlap. In the same way, tourism web pages can provide information on cultural and natural background, for instance, as well as persuade readers to visitor-specific place or region. Lastly, universities can be seen as a special type of business, where informational and advertising strategies overlap, quite in contrast, however, to teaching pages, which demonstrate culture-specific conventions and traditions.

It is clear that only a sufficient overlap of these selection criteria ensure that *ICE-Web* is compatible with the national *ICE* collections. That is why the design of the original *ICE* corpus and the adaptation to non-native corpora had to be discussed in detail (cf. Nelson 1996 and Schmied 1996 respectively).

The following diagram (Fig. 10) illustrates that *ICE-Web* is a compromise between national *ICE* corpora (including *ICE-EA*, which covers Kenya and Tanzania) and the world-wide web, which will be particularly useful for stylistic analysis and collocational research, as outlined above.

Since the WWW is skewed in various respects and does not really reflect the whole spectrum of ‘national’ usage, a systematic stratified attempt at data-collecting has to be made at least with web areas like e-commerce and e-learning. That is why the four domains were chosen: News (subcategorised into reportage, editorials and letter-to-the-editor), Business (particularly advertising and company presentation, tourism), and AcademicInstitution, i.e. the description of the institution and, AcademicLearn, i.e. learning materials. These are some typical WWW domains in most countries and thus results from the derived corpora should be based on a broad basis and internationally compatible.

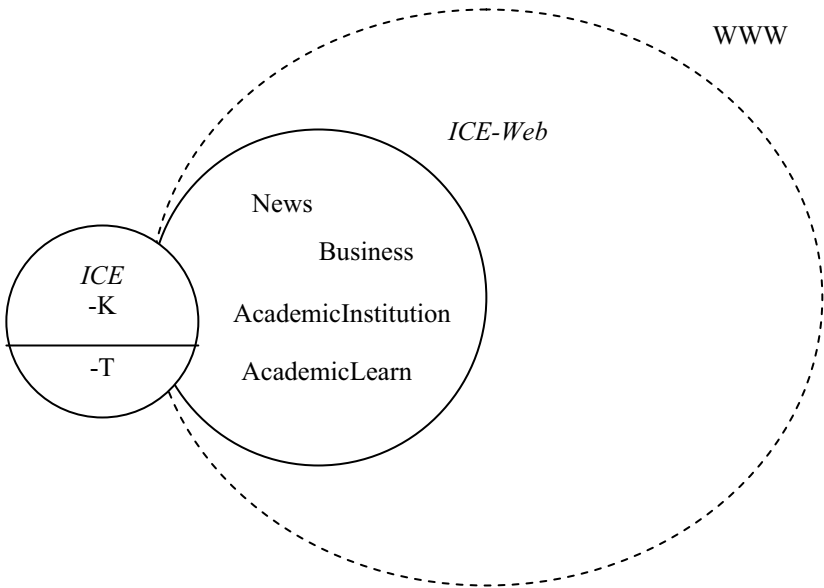


Figure 10: *ICE-Web* as a monitor corpus for *ICE-EA*.

5. Practical applications

This expanded view of *ICE-EA* can be used as a reference corpus and *ICE-Web* as a monitor corpus in all domains. Such a complex database could also be consulted by professional language-providers (media specialists, like newspaper or television reporters and web publishers) who are expected to improve texts produced for a wider or more critical readership and audience.

Thus language-conscious businessmen can compare whether the presentation of their company or particular features in it is nation-specific or international standard; the advantages and disadvantages are not easy to evaluate, since standardisation signals professionalism in the new medium, individualisation

shows independent efforts, both of which may be appreciated by customers. As is well-known nowadays, localisation is the reverse side of globalisation.

Similarly, teachers may be inclined much more to accept their students' deviations if they are also common in other ESL varieties, like India, and thus show either general creativity or irregularity on the part of Standard English (rather than their students or their own insufficiencies or interference phenomena). Testing specialists and examiners may systematize their corrections in student essays and national teaching materials. Generally, intuitions become much less important and thus corrections on the basis of statistical frequencies of usage much more objective – supplemented, of course, by a certain prestige and general acceptability in the speech community.

A reliable database of East African usage could be the basis of comparative text analyses on various levels. Since English is used as an academic language in almost all institutions of higher education in East Africa, a comparison of lecturers' notes and students' essays could indicate the gap that has to be bridged. Of course, receptive language skills have to be wider than productive ones. But appropriate specific writing classes, as opposed to more general exercises in language proficiency, might make it easier for both sides to bridge the gap satisfactorily: lectures could adapt their language through simplification, explicitness and more (even) coherence in their teaching texts; students could practise recurrent structures more specifically than using general textbooks in EAP.

Finally, the new hypermedia opportunities offer new chances for the presentation of material, which can be used in all levels of analysis:

- The presentation of pronunciation, for instance, can follow the model of the CD-ROM accompanying the current modern monolingual dictionaries for advanced learners. If texts can be presented to readers with passages highlighted in karaoke style, this makes the readers much more aware of differences, when they follow the words in a highlighted text as they are pronounced.
- Semantic analyses are the most difficult and whole branches of artificial intelligence have been working on meaning extraction, referencing and related issues. Although we are still far away from having a semantically tagged *ICE* corpus, the annotation according to word fields and collocational analyses could provide a useful starting point.
- Syntactic comparisons require intensive tagging. But even a part-of-speech (POS) tagged corpus can give a first instance of politeness differences between British and African English, for instance. The frequency and usage of modal auxiliaries like *can* or *may* indicates differences in politeness conventions that have to be analysed in detail.

Thus corpus-linguistic research could provide a substantial service to the speech community, who after all has provided the basis for analysis in the form of the multitude of texts that were needed to make the database as stratified as possible.

And the corpus-linguists in Africa, for instance, could prove that their work is of direct relevance and should be supported by administrators and politicians alike.

Notes

- 1 The corpus is freely available from the *International Corpus of English* webpage (<http://www.ucl.ac.uk/>) or with a new search tool on the project WWW pages of the *Corpus of East African English* (<http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/>). The supplementary tools *WebGrabber* and *ICE-Web* will be made available in 2006.
- 2 Just like the WWW, related corpus tools tend to change quickly; some of the screenshots in this paper are based on the older versions of *WebCorp* and *WebPhraseCount*.

References

- Fallon, H. (2004), 'Comparing world Englishes: a research guide', *World Englishes*, 23: 309-316.
- Jeffrey, C. and B. van Rooy (2004), 'Emphasizer *now* in colloquial South African English', *World Englishes*, 23: 269-280.
- Mwangi, S. (2003), *Prepositions in Kenyan English. A corpus-based study in lexico-grammatical variation*. Aachen: Shaker.
- Nelson, G. (1996), 'Markup systems', in: S. Greenbaum (ed.) *Comparing English worldwide. The International Corpus of English*. Oxford: Clarendon Press. 36-53.
- Sand, A. (2004), 'Shared morpho-syntactic features in contact varieties of English', *World Englishes*, 23: 281-298.
- Schneider, E. W. (2004), 'How to trace structural nativisation: particle verbs in world Englishes', *World Englishes*, 23: 227-249.
- Schmied, J. (1996), 'Second-language corpora', in: S. Greenbaum (ed.) *Comparing English worldwide. The International Corpus of English*. Oxford: Oxford University Press. 182-196.
- Schmied, J. (2005), 'New ways of analysing ESL on the WWW with *WebCorp* and *WebPhraseCount*', in: A. Renouf (ed.) *The changing face of corpus linguistics*. Amsterdam: Rodopi. 309-324.
- Schmied, J. (forthcoming), 'English in East Africa', in: B. B. Kachru, Y. Kachru and C. L. Nelson (eds.) *The handbook of world Englishes*. Oxford: Blackwell.

This page intentionally left blank

Transitive verb plus reflexive pronoun/personal pronoun patterns in English and Japanese: using a Japanese-English parallel corpus

Makoto Shimizu and Masaki Murata***

* Tokyo University of Science

** National Institute of Information and Communications Technology, Kyoto

Abstract

We examine the distribution of the transitive verb plus reflexive pronoun/personal pronoun patterns in English and in Japanese. We extract from a Japanese-English parallel corpus examples of the English patterns 'transitive verb followed by reflexive pronoun' and 'transitive verb followed by personal pronoun' together with their counterparts in Japanese. We estimate the distribution statistically, through a method developed by one of the authors. Then, we analyse the expressions syntactically and semantically, noting especially which types in one language correspond to which types in the other language. Certain types of verbs together with reflexive pronouns in English, for instance, are often translated into intransitive verbs in Japanese, and so on. We attempt to show that it makes more sense to focus on phrase alignments rather than word alignments when considering the correspondence between reflexive pronouns/personal pronouns in English and their counterparts in Japanese.¹

1. Introduction

The purpose of this paper is threefold. Firstly, to give a brief overview of the present situation of bilingual corpora in Japan. Secondly, to discuss the distribution of reflexive pronouns and personal pronouns in English. And thirdly, to examine transitive verb plus reflexive pronoun/personal pronoun patterns in English and their corresponding Japanese expressions.

2. Bilingual corpora in Japan

Since the 1990s a number of studies have demonstrated that bilingual parallel corpora play an important role in the fields of linguistics, corpus linguistics, computer linguistics, and translation. A number of interesting studies are to be found in Botley et al. (2000), Tognini-Bonelli (2001), and Borin (2002), to name a few.

In the case of Japanese-English parallel corpora, however, little research appears to have been done. This is due to the present situation of bilingual corpora in Japan. There are several problems concerning the compilation of Japanese-English parallel corpora such as the linguistic difference between the

two languages, copyright questions, the quality of translation, and funding. The following are a few Japanese-English parallel corpora relatively easy to obtain at present.

- *EDR Corpus*

The *EDR Corpus* is the source of the *EDR Electronic Dictionary*, a machine-tractable dictionary that catalogues the lexical knowledge of Japanese and English (the *Word Dictionary*, the *Bilingual Dictionary*, and the *Co-occurrence Dictionary*), and has unified thesaurus-like concept classifications (the *Concept Dictionary*). It has a large number of data, but it is very expensive.

- *Kansai Gaidai Corpus B*

The *Kansai Gaidai Corpus B* is a bilingual corpus of Japanese literature. The corpus is based on the CD-ROM *Shincho Bunko-no Hyakusatsu (A Hundred Books of Shincho Bunko)* and their English translations. The corpus contains a large number of texts; however, users have to purchase the original CD-ROM. See Nishimura (2002).²

- *Context Sensitive and Tagged Parallel Corpus*

The *Context Sensitive and Tagged Parallel Corpus* is a corpus of newspaper articles and editorials. The corpus is based on the *Kyodai Corpus* and its English translations and the *Penn Treebank* and its Japanese translations. The *Kyodai Corpus* is in turn based on articles and editorials published by *Mainichi Shimbun*, one of the major quality papers in Japan. The compilers employed translators to translate the *Kyodai Corpus* and the *Penn Treebank*. The *Kyodai Corpus* part contains 20,000 Japanese-English sentence pairs, and the *Penn Treebank* part contains 10,000 English-Japanese sentence pairs. Each sentence pair is tagged and aligned. See Utimoto et al. (2004).

- *Japanese-English News Article Alignment Data*

The *Japanese-English News Article Alignment Data* is a corpus of newspaper articles. It is based on articles in *Yomiuri Shimbun*, one of the major quality papers in Japan, published between 1989-2001 and its English translations, which appeared in *Daily Yomiuri*, an English daily paper published by the *Yomiuri Shimbun*. The corpus contains 150,000 Japanese sentences and 150,000 English sentences (3,570,556 words). Each Japanese-English sentence pair is aligned. The use of the corpus is free of charge. See Utiyama and Isahara (2002, 2003).

- *Gengo Shigen Kyoyuukikou*

Gengo Shigen Kyoyuukikou (GSK), a Language Resource Consortium in Japan, was founded in 1999, and is supported by the Japan Electronic Industry Development Association (JEIDA). Its purpose is to contribute to the Speech

and Language Industry in Asia. The webpage mentions some of the problems and corpora mentioned above. See the *GSK* webpage.

3. Distribution of reflexive pronouns and personal pronouns in English and Japanese

3.1 Same category hypothesis

There seems to be an assumption among some linguists that a certain type of referring expressions in one language generally corresponds to the same type in another language. Tsujimura (1996: 216), for instance, when she discusses the syntactic behaviour of Japanese reflexive pronouns, cites (1):

- (1) Taroo-ga **jibun**-o hihan-shita.
Taro-Nom self-Acc criticized
‘Taro criticized himself.’

Clearly, she is assuming that the English reflexive almost automatically corresponds to the Japanese reflexive. Let us call this view the Same Category Hypothesis (SCH).

Shimizu (1998, 2000) and Shimizu and Murata (2002a, 2002b) took issue with the SCH, discussed the correspondence between referring expressions in English and Japanese, and showed that the majority of English reflexive pronouns do not correspond to Japanese reflexive pronouns, but to what we call ‘null expressions’, while the majority of Japanese reflexive pronouns do not correspond to English reflexive pronouns, but to English personal pronouns. According to our estimation, only 10-30% of Japanese/English reflexive pronouns, depending on corpora, correspond to English/Japanese reflexive pronouns.

Shimizu and Murata (2004) focused on reflexive pronouns in the object position, which is regarded as the typical place for both English and Japanese reflexive pronouns to occur. Most previous studies of English reflexive pronouns and many of the previous studies of Japanese reflexive pronouns do not even mention other types of reflexive pronouns. We extracted from bilingual parallel corpora examples of the English patterns ‘transitive verb followed by reflexive (and preposition)’ together with their counterparts in Japanese, and analysed them syntactically and semantically. We found that English reflexive pronouns tend to correspond to several types of Japanese expressions according to the verb groups they belong to.

We have the impression that there may also be another type of SCH, relating to a single language rather than two different languages, concerning the use of English reflexive and personal pronouns. Sinclair (1990: 145-146) claims that ‘although a few verbs are typically used with reflexives, you can actually use a reflexive as the object of any transitive verb, when the meaning allows you to

do so,” and calls it “a productive feature of English”. Sinclair’s claim seems plausible since language has a creative use apart from standard grammatical rules and usages. However, we would like to raise two questions. First, is the use of a reflexive in the object position as free as Sinclair assumes? Second, what does the condition “when the meaning allows you to do so” really mean? In order to answer these questions, we examined a bilingual parallel corpus and considered the issue.

3.2 Data and method

The data we used is the *Context Sensitive and Tagged Parallel Corpus* mentioned above. We automatically extracted examples of the English patterns ‘transitive verb followed by reflexive/personal pronoun (and preposition/particle) (**Vt SELF (preposition/particle)**)’ together with their counterparts in Japanese, and manually analysed them.

We used the *Charniak Parser* to parse sentences and the *Treetagger* to stem verbs (see Charniak 1999 and Schmid 1997). We extracted sentences which matched the following patterns:

- (2) a. (VB?? ???) (NP (PRP ???))
 b. (VB?? ???) (S (NP (PRP ???)))

PRP stands for both ‘personal pronoun’ and ‘reflexive pronoun’, and S the SVOC construction such as *make it possible* exemplifies (2b).³

3.3 Results

As mentioned in 1.4, the corpus contains 150,000 Japanese-English sentence pairs. We obtained 6,614 examples, of which the verb co-occurred with reflexive pronouns in 914 examples, and with personal pronouns in 5697 examples. Thus, the ratio of reflexive pronouns was 0.1386453 and that of personal pronouns was 0.8613547. There were 941 verb types, of which 101 co-occurred only with reflexive pronouns, 684 only with personal pronouns, and 156 with both.

The three tables below show the results for representative verbs in each type. The first column shows the form of verbs, the second the occurrences of reflexive pronouns (A), the third the occurrences of personal pronouns (B), the fourth indicates the total (C), the fifth the ratio of reflexive pronouns (D), and the sixth the ratio of personal pronouns (E). The seventh and eighth columns indicate the statistically calculated improbabilities of co-occurrence of the verb with reflexive pronouns (F) and the verb with personal pronouns (G), respectively. The method for calculating the improbabilities was proposed by one of the authors of the present paper (see Murata and Isahara 2002).

Table 1: Verbs co-occurring only with reflexive pronouns.

<i>Verbs</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
distance	20	0	20	1.00000000	0.00000000	0.00000000	0.99999997
devote	14	0	14	1.00000000	0.00000000	0.00000000	0.99999413
rid	13	0	13	1.00000000	0.00000000	0.00000000	0.99998611
rehabilitate	10	0	10	1.00000000	0.00000000	0.00000000	0.99981619
reconstruct	9	0	9	1.00000000	0.00000000	0.00000000	0.99956533
rebuild	7	0	7	1.00000000	0.00000000	0.00000000	0.99757014
dissolve	6	0	6	1.00000000	0.00000000	0.00000000	0.99425597
arm	6	0	6	1.00000000	0.00000000	0.00000000	0.99425597
strengthen	5	0	5	1.00000000	0.00000000	0.00000000	0.98642300
extricate	5	0	5	1.00000000	0.00000000	0.00000000	0.98642300
brace	5	0	5	1.00000000	0.00000000	0.00000000	0.98642300
assert	5	0	5	1.00000000	0.00000000	0.00000000	0.98642300

Table 2: Verbs co-occurring only with personal pronouns.

<i>Verbs</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
use	0	89	89	0.00000000	1.00000000	0.99999630	0.00000000
sell	0	62	62	0.00000000	1.00000000	0.99983525	0.00000000
leave	0	48	48	0.00000000	1.00000000	0.99882093	0.00000000
send	0	36	36	0.00000000	1.00000000	0.99363295	0.00000000
lead	0	36	36	0.00000000	1.00000000	0.99363295	0.00000000
inform	0	32	32	0.00000000	1.00000000	0.98883062	0.00000000
urge	0	30	30	0.00000000	1.00000000	0.98520665	0.00000000
implement	0	29	29	0.00000000	1.00000000	0.98297515	0.00000000
accept	0	29	29	0.00000000	1.00000000	0.98297515	0.00000000
want	0	28	28	0.00000000	1.00000000	0.98040710	0.00000000
order	0	28	28	0.00000000	1.00000000	0.98040710	0.00000000
offer	0	28	28	0.00000000	1.00000000	0.98040710	0.00000000
pay	0	27	27	0.00000000	1.00000000	0.97745174	0.00000000
encourage	0	27	27	0.00000000	1.00000000	0.97745174	0.00000000

Table 3: Verbs co-occurring with both reflexive and personal pronouns.

<i>Verbs</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
turn	10	11	21	0.47619048	0.52380952	0.00000000	0.00000000
protect	16	18	34	0.47058824	0.52941176	0.00000000	0.00000000
distinguish	5	6	11	0.45454545	0.54545455	0.00000000	0.00000000
convert	5	6	11	0.45454545	0.54545455	0.00000000	0.00000000
open	3	4	7	0.42857143	0.57142857	0.00000000	0.00000000
enjoy	3	4	7	0.42857143	0.57142857	0.00000000	0.00000000
set	7	10	17	0.41176471	0.58823529	0.00000000	0.00000000
find	87	127	215	0.40465116	0.59534884	0.00000000	0.00000000
kill	14	21	35	0.40000000	0.60000000	0.00000000	0.00000000

restore	2	3	5	0.40000000	0.60000000	0.00000000	0.00000000
conduct	2	3	5	0.40000000	0.60000000	0.00000000	0.00000000
change	2	3	5	0.40000000	0.60000000	0.00000000	0.00000000
appoint	2	3	5	0.40000000	0.60000000	0.00000000	0.00000000
relieve	3	5	8	0.37500000	0.62500000	0.00000000	0.00000000
isolate	3	5	8	0.37500000	0.62500000	0.00000000	0.00000000
control	4	7	11	0.36363636	0.63636364	0.00000000	0.00000000

We will examine in detail some of the verbs and reflexive/personal pronouns in the next section.

4. Transitive verb plus reflexive pronoun/personal pronoun

We analysed the verb and reflexive/personal pronoun combinations syntactically and semantically, noting especially which types in one language correspond to which types in the other language.

4.1 Patterns

It seems that a verb prefers a certain pattern when it co-occurs with reflexive pronouns, and prefers another pattern when it co-occurs with personal pronouns. Famous examples are *find*, *see*, and *consider*, which Barlow (1996) calls distancing or viewing verbs. Barlow claims that this type of verb occurs most frequently with reflexive pronouns, and the typical pattern in which *find oneself* occurs is **Vt SELF participle/PP**. Fukaya (2002) makes several interesting observations on the reflexive pronouns occurring in this pattern. In *Daily Yomiuri*, *find* co-occurs 87 times with reflexive pronouns, and 79.3% of these occurrences are in the pattern **Vt SELF participle/PP**. Two examples together with their Japanese counterparts are given in (3).

- (3) a. And so Algeria finds **itself** poised on the brink of civil war.
 Koshite Arujeria-wa naisen-no setogiwa-ni ok-areru koto-ni nat'ta.
 Thus Algeria-TOP civil-war-GENverge-LOC be-put thing-LOC
 become-PAST
 "Thus Algeria reached the state of being on the verge of civil war."
 b. Kubo has found **himself** in an uncomfortable position.
 Kubo zosho-no tachiba-mo kimyona koto-ni nat'te-shimau.
 Kubo Minister-of-Finance-GEN situation-COM strange thing-LOC
 become-PERF
 "Finance Minister Kubo's situation has become strange."

In contrast, *find* co-occurs 128 times with personal pronouns, and 89.0% of occurrences are in the pattern **Vt PRON adj**. The majority of the personal pronouns are so-called 'dummy' or 'anticipatory' objects, and most of the

adjectives are related to judgement such as *easy*, *hard*, *difficult*, and *impossible*, as shown in (4).

- (4) All this has made it clear that the two nations will find **it** impossible to conclude the long-awaited peace treaty by the end of the year.
 Kono kekka, nen'nai-no heiwa-jyoyaku-teiketsu-wa, genjitsu-mondai-toshite fukano-dearu koto-ga hakkiri-shi-ta.
 Impossible-be-PRES thing-NOM clear-become-PAST (related parts only)
 “(that peace treaty) is impossible became clear”

Distance co-occurs with reflexive pronouns 18 times, but never co-occurs with personal pronouns. It is always used in the pattern *distance SELF from*.

- (5) North Korean authorities distance **themselves** from reform because they fear a collapse of the system.
 Kitachosen-shidoso-wa touchi-taisei-ga shometsu-suru osore-kara,
 kaikaku-kara kao-o somukete-iru.
 reformation-SOURCE face-ACC turn-away-PROG (related parts only)
 “(North Korea) is turning its face away from the reformation”

The same holds true for *devote* (14 times, all *devote SELF to*) and *rid* (13 times, all *rid SELF of*). The verb *concern* co-occurs 3 times with reflexive pronouns, and 7 times with personal pronouns. When co-occurring with reflexive pronouns, the pattern is always *concern SELF with*. *Allow* co-occurs 3 times with reflexive pronouns, and 94 times with personal pronouns. When co-occurring with reflexive pronouns, the pattern is always *allow SELF to be PP*. It seems that this pattern has a negative connotation.

- (6) It was a fatal mistake by Yasser Arafat, chairman of the Palestine Liberation Organization, to allow **himself** to be duped by Saddam.
 Paresuchina-kaihou-kikou-no Arahato gicho-wa,
 kono kuchiguruma-ni not'te Iraku-shiji-toiu chimeitekina handanmisu-o okashita.
 this cajolery-LOC get-on-PRES (related parts only)
 “Arafat was cajoled into (backing up Iraq)”

Commit co-occurs 36 times with reflexive pronouns, of which 32 cases comprise the pattern *Vt SELF to*, 2 cases *Vt SELF in*, and 2 cases *Vt SELF*. NPs come after the pattern *Vt SELF to/in* in 17 cases, gerunds in 11 cases, and bare infinitives in 6 cases. The verb co-occurs once with personal pronouns. An example together with its Japanese counterpart is given in (7).

- (7) Japan, for its part, committed **itself** to financial support totaling 64 billion dollars to help revive East Asian countries hit by the crisis.
 Nihon-mo Miyazawa-o hajime-toshite 640 oku doru-ni noboru
 shikin-teki shien-o komit'to-shita.
 financial support – ACC commit-do-PAST (related parts only)
 “(Japan) promised financial support”

Note that the Japanese counterpart of the English verb *commit* is *komit'to-suru*, a loan word from *commit*. Obviously, the writer of the Japanese sentence felt inclined to use the verb in preference to other non-loan words. It is interesting that the pattern is not *jibun-o -ni komit'to-saseru* but *-ni komit'to-suru*, that is, not **Vt JIBUN**, but **Vi**.

Identify co-occurs equally with reflexive pronouns (13 times) and personal pronouns (12 times). However, when it occurs in the pattern **Vt with**, the slot is filled only with reflexive pronouns.

- (8) Poland, the Czech Republic and Hungary have, in effect, identified **themselves** with the West, which is only reasonable considering the religious and cultural heritage and traditions they share with Western nations.
 Porando, Cheko, Hangarii-wa shukyoteki-nimo bunkateki-nimo
 seiou-tono kyoutsuten-ga ooku, ganrai, 'nishi'-eno kizoku-ishiki-ga tsuyoi.
 the-West-COM something-in-common a-lot (related parts only)
 “there are a lot of things in common between the West and (Poland, the Czech Republic and Hungary)”

The genre of journalism might affect the use of certain groups of verbs such as *arm*, *dissolve*, and *liquidate*. These verbs are concerned with the military, politics, and economy, topics covered by journalism on a daily basis. For instance, *arm* co-occurs with reflexive pronouns 6 times, but never co-occurs with personal pronouns. In 4 cases, the pattern is *arm SELF with*. The corresponding Japanese expression is *buso-suru* in 4 cases.

- (9) As a by-product of the North Korea issue, predictions have been made in various parts of the world that Japan will move to arm **itself** with nuclear weapons in the future.
 Kitachosen-no kaku-giwaku-no fukujiteki shosan-toshite, nihon-wa shorai
 kaku-busou-ni susumude-arou-toiu yosoku-o sekai-no kakuchi-ni unda.
 Nuclear-armament-LOC proceed-may-that prediction-ACC
 (related parts only)
 “prediction that (Japan) may proceed to armament with nuclear weapons”

Francis et al. (1996) list *disarm SELF with*, but not *arm SELF with*. The *Longman Dictionary of Contemporary English* gives examples of *arm* co-

occurring with reflexive pronouns, but has no comment to make on the use of reflexive pronouns.

It is interesting to note that *rebuild* (7 times), *reconstruct* (9 times), *rehabilitate* (10 times), and *reorganise* (7 times) tend to co-occur with reflexive pronouns in the corpus. The first three never co-occur with personal pronouns, and the last co-occurs only once with personal pronouns.

- (10) The government's reform priority program plans to establish a fund for corporate restructuring, designed to help business corporations rebuild **themselves**.

Seifu-no kaikaku puroguramu-niwa, kigyo-no saiken-o hakaru
 corporation-GEN reconstruction-ACC try (related parts only)
 'kigyo saiken huando'-no souken-ga morikom-are-ta.
 "(The government) try corporations' reconstruction"

The reason may be the bound morpheme *re-* because *build* and *organise* co-occur with personal pronouns twice, and *construct* once, but they do not co-occur with reflexive pronouns. Most of the corresponding Japanese expressions also have the bound morpheme *sai-*, which means 'again'.

Control SELF is somewhat of an exception in our data because in the Japanese counterparts there appear *jibun* and *jiko*, free morphemes meaning 'self' or *ji-*, bound morphemes meaning 'self'.

- (11) All this would cause anyone to presume that the suspect may have been spoiled by his family as a child and grown up without learning how to control **himself**.

Katei-de amayakas-are, gaman-ya jiko-kontoruru-o shira-nai-mama-ni
 Patience-and self-control-ACC know-not-remain (related parts only)
 sodat'ta otoko-no sugata-ga ukabu.
 "remaining ignorant of patience and self-control"

4.2 Five strategies

We have found at least five strategies to make Japanese expressions correspond to **Vt SELF** in English. Sometimes more than one strategy can be used for a single verb. Even though the choice of a strategy is, in such a case, at the translator's discretion, it seems that there exist some tendencies. The first strategy is to intransitivise the transitive verb, namely, just delete the reflexive pronoun, especially when the verb is ergative.

- (12) This is a manifestation of the Japanese economy's all-out efforts to adjust **itself** to the high value of the yen and global structural changes that followed the end of the Cold War.
 Daga, kore-wa nihon-keizai-ga hageshii endaka-ya reisen-go-no sekai-keizai-no henka-ni kenmeini taiou-suru sugata-o shimesu-mono-dearu.
 global-economy-GEN change-LOC hard adjustment-do (related parts only)
 "(try) hard to adjust to the change of world economy"

The second strategy is to paraphrase the verb, often with bound morphemes.⁴

- (13) However, having watched as U.S. banks expanded through mergers, developed new financial products one after another, began dealing in derivatives and other instruments and took the lead globally through massive investment in information technology, Deutsche Bank decided to transform **itself**.
 Shikashi, Amerika-no ginko-ga gap'pei-de kibo-o kakudai-suru ip'po, deribatibu-nadode tsugitsugi-to atarashii kinyu-shohin-o kaihatsu-shi, kyogaku-no toushi-o susumete jyouhou-gijutsu-de sekai-o riido-ruru-no-o mite, henshin-o ketsudan-shita.
 (related parts only) transformation-ACC decision-do-PAST
 "(Deutsche Bank) made a decision on (its) transformation"

In the Japanese sentence, the word *henshin* appears. The morpheme *hen-* means 'change', and *-shin* means 'body'.

The third strategy is to change the reflexive pronoun into a noun, especially when the verb is a 'groom', 'hurt', 'dress' verb (Levin 1995) or a 'logophoric' verb (Kuno 1987). These nouns tend to be the whole or a part of the human body, clothes, speech, idea, feeling, health, looks, figure, behaviour, and so on:⁵

- (14) At this juncture in time, when we face difficult issues, both at home and abroad, I intend to brace **myself** to pour every effort into this endeavor.
 Naigai-ni kon'nan-na kadai-o kakaeru kon'nichi kokoro-o hikishime,
 (related parts only) Heart-ACC brace
 zenryoku-de torikunde-mairi-masu.
 "(I will) brace (my) heart and tackle (the issues) with all (my) force"

The fourth strategy is to paraphrase the **Vt SELF** into an expression with the bound morpheme *ji-* or *jiko-*, which means 'self'.

- (15) Reflecting the harsh desert environment in which he was raised, his abrasive manner of expressing **himself** often angers people, although those

close to him brush it off as an unfortunate mannerism that should not be taken too seriously.

Jibun-ga sodat-ta sabaku-no kibishii kankyo-o hanei-suru-ka-noyouna
kare-no mimizawarina jiko-hyougen-youshiki-wa, shibashiba hito-o
okoraseru.

He-GEN harsh self-expression-manner-TOP (related parts only)
“his harsh manner of self-expression”

The fifth strategy is, on the last resort, to use a reflexive pronoun.

- (16) He said he would have blamed **himself** if he had lost the car in an accident, but because it had been stolen, he was angry.
“Jiko-dat'tara jibun-o semer-areru-ga --- “ to ikidooru.
myself-ACC blame-can-but (related parts only)
“I could (have) blame(ed) myself, but ---”

Example (16), however, seems to us a kind of translationese, bearing in mind Johansson and Hofland's (1994: 26) definition of translationese as “features of the translated text more characteristic of the source language than the language the sentence is being translated into”.

4.3 English reflexive/personal pronouns corresponding to *jibun*

Although the main topic of discussion in this paper has been English reflexive pronouns, we did a brief research on the Japanese reflexive pronoun *jibun*. We automatically extracted Japanese sentences which contained *jibun*, checked the corresponding English sentences as to whether they contained reflexive or personal pronouns, and counted them when they did.

As mentioned in 2.2, 914 reflexive pronouns and 5,697 personal pronouns occur in the English sentences. Out of 914 reflexive pronouns, 38 occur in the corresponding Japanese sentences, and out of 5,697 personal pronouns, 77 occur in the corresponding Japanese sentences. Thus, the Japanese reflexive pronoun *jibun* is likely to be translated into English personal pronouns more often than English reflexive pronouns. This result coincides with that of our previous research.

5. Conclusion

We extracted examples of the English patterns **Vt SELF/PRON** together with their counterparts in Japanese, and analysed them syntactically and semantically. It is difficult to give definite answers to the questions raised in 2.0 since the number of examples is rather small because of the size of the corpus.

With regard to the first question, we would tentatively suggest, however, that the use of a reflexive in the object position is not as free as Sinclair assumes.⁶ We agree with Tognini-Bonelli (2001: 101) that “we have to abandon the fiction

that each word is some kind of independent selection, and accept that the choice patterns of words in text can create new, large and complex units of meaning.”

As for the second question, although we pointed out a few semantic factors, we must admit that we need further investigation.

Notes

- 1 We would like to express our gratitude to Mira Ariel for her valuable comments on the previous version of this paper. Any remaining errors are, needless to say, our own responsibility.
- 2 We would like to thank Professor Nishimura at Kansai Gaidai University for letting us use *Kansai Gaidai Corpus B*.
- 3 We used the two rules to shorten the time needed for data extraction. Admittedly, the rules are too simple to exclude errors completely. We believe, however, that the general tendency should be similar.
- 4 Mira Ariel (personal communication) commented the second strategy and the third strategy may essentially be the same. Although examples (13) and (14) look similar semantically, the difference between them is the morphological status. While *-shin* in (13) is a bound morpheme, *kokoro* in (14) is a single lexical word. In other words, **Vt SELF** in the English sentence corresponds to a single word in the Japanese sentence in (13), and to several words in (14). It might be possible to claim that more lexicalised expressions are used in the second strategy.
- 5 Mira Ariel (personal communication) calls this a proto-reflexive pronoun strategy, and points out that a number of languages in the world use this strategy.
- 6 In Mira Ariel’s opinion, when Sinclair claims a productive feature of English, what he means by *reflexives* is bound anaphora cases. We are not certain whether Sinclair assumes that there are two kinds of reflexives. Even if he does, we argue that he needs to present the criterion to distinguish the two.

References

- Barlow, M. (1996), ‘Corpora for theory and practice’, *International journal of corpus linguistics*, 1: 1-37.
- Borin, L. (ed.) (2002), *Parallel corpora, parallel worlds*. Amsterdam: Rodopi.
- Botley, S. P., A. M. McEnery and A. Wilson (eds.) (2000), *Multilingual corpora in teaching and research*. Amsterdam: Rodopi.

- Charniak, E. (1999), *A maximum-entropy-inspired parser*, Technical Report CS-99-12, Brown University [online] Available from: <http://acl.ldc.upenn.edu/A/A00/A00-2018.pdf>.
- Francis, G., S. Hunston and E. Manning (1996), *Verbs: patterns and practice*. London: Harper Collins.
- Fukaya, T. (2002) 'On viewing reflexives in the Bank of English: their distribution and function', in: T. Saito, J. Nakamura and S. Yamazaki (eds.) *English corpus linguistics in Japan*. Amsterdam: Rodopi. 77-91.
- Johansson, S. and K. Hofland (1994), 'Towards an English-Norwegian parallel corpus', in: U. Fries, G. Tottie and P. Schneider (eds.) *Creating and using English language corpora*. Amsterdam: Rodopi. 25-37.
- Kuno, S. (1987), *Functional syntax: anaphora, discourse and empathy*. Chicago: University of Chicago Press.
- Levin, B. (1995), *English verb classes and alternations: a preliminary investigation*. Chicago: University of Chicago Press.
- Murata, M. and H. Isahara (2002), 'Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples based on positive examples', *IEICE transactions on information and systems*, vol. E85-D, 9: 1416-1424.
- Nishimura, T. (2002), '*Kansai Gaidai Corpus B-no Gaiyou* (Compilation of *Kansai Gaidai Corpus B*)', *English Corpus Studies*, 9: 37-43. (In Japanese).
- Sinclair, J. (ed.) (1990), *Collins Cobuild English grammar*. London: Harper Collins.
- Shimizu, M. (1998), 'Nichieigo-no alignment mondai nitsuite: saikikei-o chushin-ni (The alignment problem of the Japanese and English reflexives)', in: *Papers from the fifteenth national conference of the English linguistic society*. 171-180. (In Japanese).
- Shimizu, M. (2000), 'Nichieigo-kan niokeru saikikei-no honyaku mondai (The translation problem of the Japanese and English reflexives)', *Bulletin of Tokyo University of Science (Dept of Humanities)*, 32: 29-39. (In Japanese).
- Shimizu, M. and M. Murata (2002a), 'Parallel corpus-o mochiita nichieigo saikikei-no bunseki (On the English and Japanese reflexives and their translations)', *English corpus studies*, 9: 17-34. (In Japanese).
- Shimizu, M. and M. Murata (2002b), 'Nihongo-no kuuhyogen nitaisuru saikikei nitsuite (On the Vt+-self/-selves construction in English)', *Bulletin of Tokyo University of Science (Dept of Humanities)*, 34: 29-42. (In Japanese).
- Shimizu, M. and M. Murata (2004), 'Patterns with transitive verb and reflexive in English and their counterparts in Japanese: a bilingual pattern grammar approach', in: J. Nakamura, N. Inoue and T. Tabata (eds.) *English corpora under Japanese eyes: JAECS anthology commemorating its 10th anniversary*. Amsterdam: Rodopi. 71-91.

- Schmid, H. (1997), 'Probabilistic part-of-speech tagging using decision trees', in: D. Jones and H. Somers (eds.) *New methods in language processing*. London: UCL Press.
- Tognini-Bonelli, E. (2001), *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Tsujimura, N. (1996), *An introduction to Japanese linguistics*. Malden: Blackwell.
- Uchimoto, K., K. Sudo, M. Murata, S. Sekine and H. Isahara (2004), 'Context sensitive and tagged parallel corpora', in: *Proceedings of the 10th annual meeting of the association for natural language processing*. Osaka: The Association for Natural Language Processing. 592-595.
- Utiyama, M. and H. Isahara (2002), 'Nichiei shinbun kiji-no taiouzuke to sono hyoka (Alignment of Japanese-English news articles and sentences)', in: *Working group of natural language technical reports, 2002-NL-151*. Tokyo: Information Processing Society of Japan. 15-22. (In Japanese).
- Utiyama, M. and H. Isahara (2003), 'Reliable measures for aligning Japanese', in: *Proceedings of ACL-2003*. East Stroudsburg: The Association for Computational Linguistics. 72-79.

Webpages

- EDR Corpus* <http://www.ijjnet.or.jp/edr/index.html>
- GSK* <http://www.gsk.or.jp>
- Japanese-English News Article Alignment Data*
<http://www2.crl.go.jp/jt/a132/members/mutiyama/corpora.html>
- JEITA* <http://www.jeita.or.jp/english/index.htm>
- Shinchosha* <http://www.shinchosha.co.jp/index.html>

Dictionaries

- Longman dictionary of contemporary English* (3rd ed.). Harlow: Longman. [CD-ROM].

The retrieval of false anglicisms in newspaper texts

Cristiano Furiassi* and Knut Hofland**

* University of Turin

** Aksis, University of Bergen

Abstract

The present article is the description of a project aimed at building a specialized corpus of Italian newspaper texts and at developing a computational technique to retrieve new false anglicisms from it. Texts were collected along a ten-month span from three Italian newspapers: La Stampa, La Repubblica, and Il Corriere della Sera. The size of the corpus is about 20 million tokens and approximately 230,000 types. The system was automatically updated on a daily basis and a list of words was obtained at the end of the collection period. This procedure originated a refined word list in which false anglicisms were searched. Along with computational techniques, careful manual scanning proved to be indispensable to extract new false anglicisms. The corpus is available for future work and may be exploited not only to find false anglicisms but also to retrieve anglicisms, neologisms, and to analyse lexical features of Italian newspaper language.

1. Introduction

The present article describes a project aimed at building a specialized corpus of Italian newspaper texts to be used for retrieving new false anglicisms. Previous research on false anglicisms based on lexicographic resources has already provided a ‘framework’ for the identification of these English-looking words, as well as an inventory of false anglicisms in Italian (see Furiassi 2003). This is briefly summarised in the first part of the article.

In order to find ‘new’ false anglicisms, the language of newspapers is most suitable since it is representative of a wide range of registers and is highly receptive and open towards neologisms, loanwords and linguistic creativity in general. Foreign words, anglicisms and false anglicisms in particular, are very often used for their positive connotation, their strategically communicative features, and intrinsic “stile brillante” (Marello 1996: 32). The connotative meaning associated to English or English-looking words is perhaps the main reason why they are used in newspaper articles and especially in eye-catching headlines.

The press plays a very important role as a primary source for the introduction of false anglicisms in the Italian language. The fact that false anglicisms are often highlighted and explained by glosses is likely to fix them in both the passive and the active lexical competence of Italian newspaper readers.¹ Written texts give way to actively mediated and formally motivated vocabulary acquisition, thus being of particular interest for the study of false anglicisms and

granting them more chances to survive in the tangle of newly introduced borrowings and coinages.

The second part of the article describes the compilation of a corpus of Italian newspaper texts and the computational technique which was implemented in order to isolate new false anglicisms. Computer corpora are extremely useful to study loanwords in general and more specifically false anglicisms (Bevitori 2002: 64). Apart from offering a large searchable linguistic database, electronic corpora enable the researcher to evaluate whether a certain entry is to be considered a false anglicism or not. In addition, the most common collocations for each false anglicism help discriminate between alternative senses by highlighting the typical contexts of occurrence in which false anglicisms tend to appear. Finally, examples taken from corpora also show some grammatical features of false anglicisms.

2. A definition of false anglicisms

The average Italian speaker does not seem to be aware of the fact that many English-looking and/or English-sounding words are not English at all (Pulcini 2002: 163-165). These words are usually referred to as 'pseudo-anglicisms' or 'false anglicisms'.

False anglicisms are autonomous creations of the Italian language that formally resemble English words but actually do not belong to the English language, even though they are recognised as authentic English elements by Italian speakers (e.g. *autostop* instead of *hitch-hiking*). The fact that false anglicisms are made with real English words implies that the Italian speaker who uses them must have at least some knowledge of the English language.

There are also false anglicisms that are proper English words but are used in Italian with totally different meanings (e.g. *smoking* instead of *tuxedo* or *dinner jacket*). At any rate, false anglicisms do not undergo any kind of orthographic or morphological adaptation, that is, they are not integrated into the orthographic structure of Italian.²

As to morphology, derivatives such as *filmino* should be considered the output of the assimilation process of borrowings (i.e. adapted anglicisms) rather than false anglicisms, since the resulting form has no equivalent in English. Hence, the adding of suffixes is a linguistic process which adjusts the borrowing to the structural patterns of the Italian language. False anglicisms may only be adapted in pronunciation in order to comply with the phonological system of Italian.

False anglicisms are either formally or semantically different from the original English words from which they are supposed to derive, so that both an English native speaker, proficient in Italian, and an Italian native speaker, proficient in English, would recognize them in spoken and written registers. (Furiassi 2003: 123).

Finally, the status of false anglicisms does not always imply that such items necessarily lack a native lexical model; false anglicisms might sometimes mirror real English words.

3. A classification of false anglicisms

False anglicisms originate from various linguistic processes and it is often difficult to determine their origin. Three fundamental types of false anglicisms may be distinguished and classified: autonomous compounds, compound ellipses, and semantic shifts (Furiassi 2003: 124-125).

3.1 Autonomous compounds

False anglicisms having the form of autonomous compounds are non-English compounds with two elements that can be separately found in English, whose composite form, however, is a genuine Italian product independent of any specific model. This leads to the coinage of brand-new false anglicisms by means of real English words. Such compounds are not used in native varieties of standard English.³

A typical example of an autonomous compound is the word *camera car*, which is not used in English but is actually composed of two authentic English elements (i.e. *camera* and *car*). Actually, the correspondent English equivalent is *on-board camera*.⁴

3.2 Compound ellipses

Even though the ellipsis of compounds is quite a common word-formation process in both English and Italian, certain compound ellipses of English words are characteristic of the Italian language (Iamartino 2001: 121). For instance, the compound ellipsis *skate*, for *skateboard*, is accepted in English. Conversely, the word *basket* does not function as the elliptical form of *basketball*. In fact, the false anglicism *basket* would not be intuitively associated to *basketball* by a native speaker of English.

Therefore, false anglicisms originated from the ellipsis of an English compound can be formally found in English, though with a different meaning. After becoming lexical units independent of the English compounds from which they derive, such elliptical forms may hinder the full comprehension of the word.

Moreover, non-English compound ellipses derive their meanings from the entire English compounds which have been truncated. In English, which is a pre-modifying language, compound ellipses are more likely to occur by eliminating the word on the left (e.g. *night club* may be shortened into *club*). Conversely, in Italian, which is a post-modifying language, the head usually comes first and is followed by the modifying element. Therefore it is the right-hand element which is normally deleted (e.g. *night club* is shortened into *night*).

3.3 Semantic shifts

Semantic shifts involve a process of meaning change. A false anglicism derived from a semantic shift is a word that may be encountered in English but that takes on a new meaning in Italian depending on the context in which it appears. The meaning given to such items strikes the ordinary English speaker as counterfeit. In Italian, for example, the word *mister* refers to the trainer of a sports team. The appropriate English equivalent would actually be either *coach* or *trainer*.

Even though they are formally identical in both languages, false anglicisms originated from a semantic shift are words that have kept a genuine English form whose meaning, however, significantly departs from the English homograph. Therefore, the acceptable degree of semantic difference which would allow certain items to be labelled as false anglicisms is incomprehensibility, since it is evident that the attribution of new meanings to Italian false anglicisms, which have corresponding English homographs, leads to ambiguity.

4. The *HF Corpus*

In order to retrieve and study false anglicisms in Italian, a corpus of newspaper language was compiled and analyzed.⁵ The corpus, which was created *ad hoc* for the analysis of false anglicisms, was named *HF* after the surname initials of its compilers. The acronym also recalls the name of a building at the university campus where the project started. The *HF Corpus* was compiled at the *Department of Culture, Language, and Information Technology (Aksis)* of the *University of Bergen, Norway*. The main goal of the *HF Corpus* was the compilation of an up-to-date database of newspaper articles from which false anglicisms may be extracted. The corpus, which is accessible on-line exclusively for research purposes, will be available for future work not only to find false anglicisms but also to retrieve anglicisms, neologisms, and other lexical features of Italian newspaper language.

The data shown below refer to the amount of text gathered along a ten-month span (August 2003 – May 2004) from the web sites of three Italian newspapers: *La Stampa*, *La Repubblica*, and *Il Corriere della Sera*. The system automatically updates on a daily basis and a list of words is continuously produced in order to look for possible new false anglicisms at the end of the collection period.

The method used to select articles and include them in the *HF Corpus* employed enhanced *Unix* scripts combined with *w3mir* software, which was used to retrieve *HTML* texts from the newspapers considered (see Hofland 2000).⁶ Captions, menus, and links to other pages were deleted and the size of the texts collected amounts to about 24.34 million tokens (about 13.44 million from *La Repubblica*, about 6.51 million from *Il Corriere della Sera*, and about 4.39 million from *La Stampa*) and 384,414 types. After eliminating proper nouns – by removing upper-case words – and numbers, the total number of tokens amounts to about 19.47 million and types are 232,001.

The tangible achievement is a small yet up-to-date corpus of Italian newspaper language. The software used to search the corpus is based on *Corpus Workbench* (CWB) and each newspaper may also be searched separately (Figure 1).

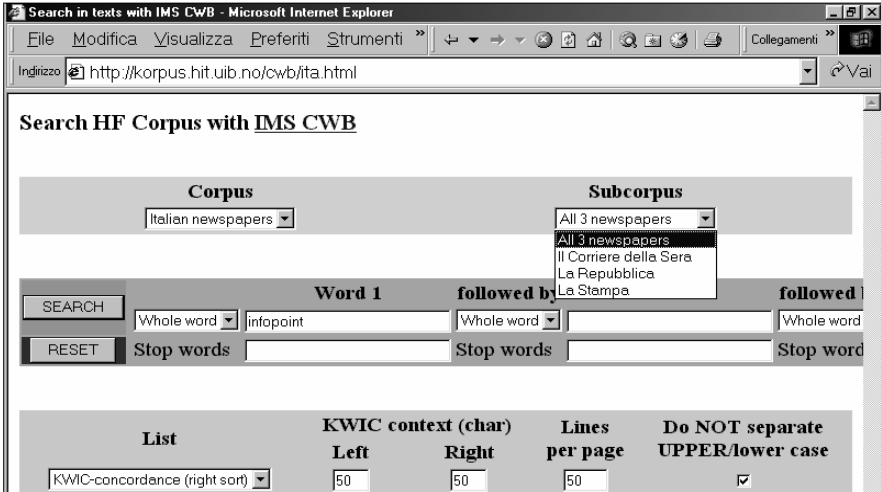


Figure 1: The *HF Corpus* search window.

By means of the search window, all the occurrences of the searched item (e.g. *infopoint*) are displayed and numbered. The date on which a certain item appears and the source newspapers are indicated on the left of the concordance line (Figure 2).

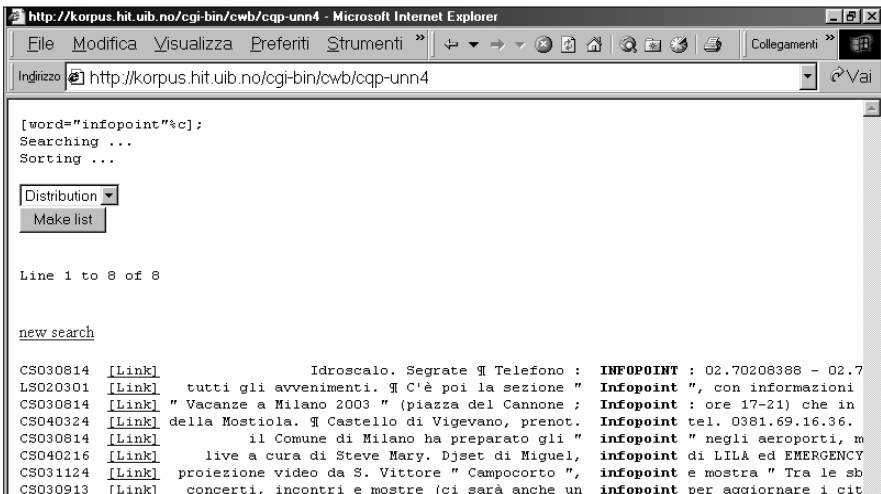


Figure 2: Sample search.

In addition, by clicking on the ‘Link’ button at the beginning of each concordance line, the software shows the original web page from which the false anglicism was taken. Finally, by clicking on the ‘Make list’ button, the user may observe the quantitative distribution of the item in the various newspapers considered, i.e. CS: *Corriere della Sera*, LR: *La Repubblica*, and LS: *La Stampa* (Figure 3).

abs. freq	rel. per 100 mill.	code
7	108	CS
1	23	LS
8	33	total

Figure 3: Quantitative distribution.

5. The retrieval of false anglicisms in the *HF Corpus*

The *HF Corpus* originated a list of 232,001 items (i.e. types). Subsequently, the list was refined according to various automatic procedures and by means of computational tools and techniques in order to isolate false anglicisms.

The selection procedure may be divided into two different though connected parts, both aiming at the retrieval of false anglicisms from the *HF* word list. The first method exploits the intersection of selected word lists with the one obtained from the corpus. The second method combines *n*-gram statistics and word list intersection. At a later stage, the two methods were merged (Figure 4).

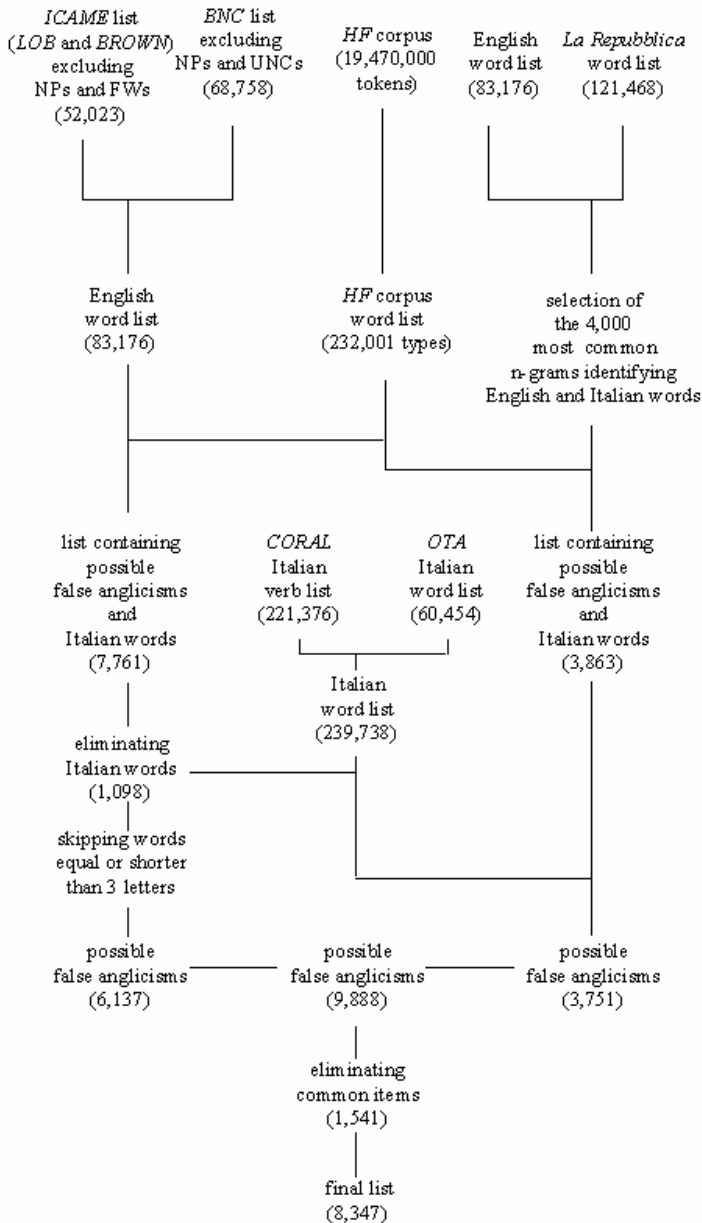


Figure 4: The retrieval of false anglicisms in the *HF Corpus*.

5.1 Intersecting word lists

The intersection of word lists allowed the selection of real English words and the exclusion of real Italian words from the *HF* word list in order to isolate potential false anglicisms from the outcoming list. The *HF* list was first intersected with a list acquired from a combined *ICAME-BNC* list of 83,176 word forms. The *ICAME* list, which amounts to 52,023 items, was obtained by excluding words tagged as NP (i.e. proper nouns) and FW (i.e. foreign words) from a list of words included in the POS-tagged versions of the *London-Oslo-Bergen (LOB) Corpus* and the *BROWN Corpus*, which are two of the several sub-corpora belonging to the *International Computer Archive of Modern and Medieval English (ICAME)*. The *BNC* list, which contains 68,758 items, was obtained by simultaneously excluding words tagged as NP (i.e. proper nouns) and UNC (i.e. unclassified words) and keeping only words occurring in at least six different texts from the *British National Corpus*. After eliminating strings equal or longer than three graphic words in order to exclude quotations, this step allowed the selection of 7,761 English words which were included in the *HF* list.

Secondly, in order to eliminate Italian words, a combined *OTA-CORAL* list of 239,738 word forms was intersected with the *HF* list. The list selected from the *Oxford Text Archive (OTA)* includes 60,454 Italian words and the lemmatised list of Italian verbs obtained through the *Corpora e Apprendimento Linguistico (CORAL)* program includes 221,376 items.⁷ This step led to the elimination of 1,098 Italian words which were originally included in the *HF* word list.

This procedure generated a provisional list of about 6,663 words which still included some ‘noise’, i.e. abbreviations, acronyms, and prepositions, which was then submitted to further automatic skimming. In order to exclude abbreviations, acronyms, and prepositions, words shorter than or equal to three orthographic characters were deleted. This latter device did not affect the search since no false anglicism equal or shorter than three letters exists, as shown by previous research based on lexicographic resources (see Furiassi 2003). Finally, a list of 6,137 items, which is very likely to contain new false anglicisms, was obtained (Figure 5).

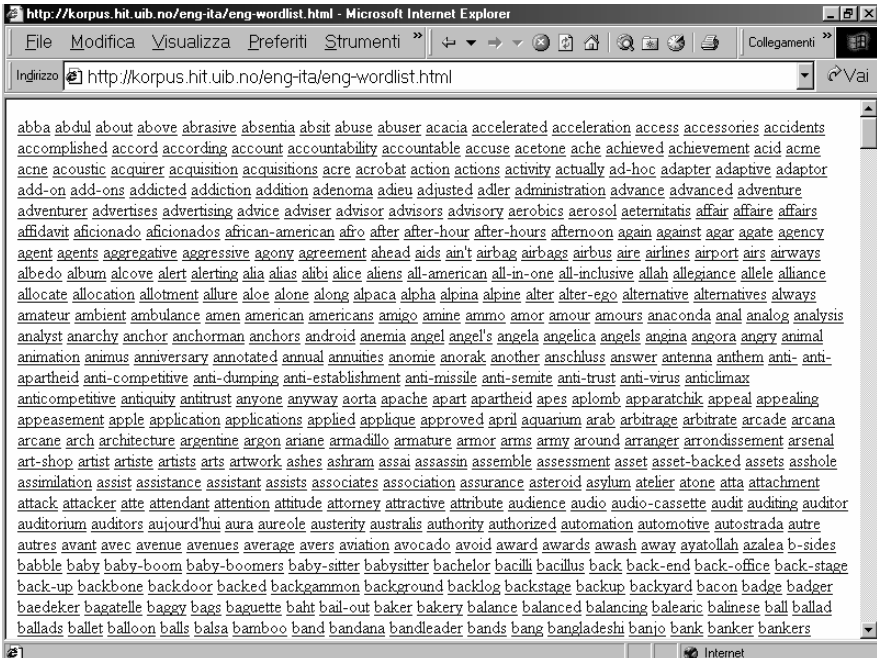


Figure 5: The list obtained from the intersection of word lists.

5.2 Using *n*-gram statistics

N-grams are recurrent combinations of items of various kinds (e.g. graphemes, morphemes, words, phrases, and sentences) which may be customised according to the user's needs. *N*-grams are useful for several linguistic functions – including finding collocations for machine translation and automatic tagging of texts – and provide insights into language usage. For the aims of the present work, *n*-grams of characters were used to recognise the patterns of a language (i.e. English) and to identify items belonging to it according to the probabilities that certain grapheme-combinations have in that language.⁸

The following table shows a selection of the first 50 *n*-grams out of the 4,000 most frequent (Table 1). The apex symbol ‘^’ indicates beginning of word or end of word, whether it proceeds or follows the string of letters. For instance, there are 19,574 Italian words ending with the letter ‘o’ and 210 English words starting with ‘ri’ in the *HF Corpus*. Strings with no apex symbol show that a certain sequence of graphic characters is found within a certain word.

Table 1: The 50 most frequent *n*-grams.

	<i>n-gram frequency (English)</i>	<i>n-gram frequency (Italian)</i>	<i>n-gram sequence</i>
1	314	19,574	o [^]
2	93	16,224	i [^]
3	12,245	52	s [^]
4	3,872	9,735	ra
5	449	9,083	a [^]
6	2,365	8,271	ta
7	7,096	54	d [^]
8	7,068	1,413	ed
9	521	6,890	av
10	1,880	6,888	as
11	1,427	6,645	ia
12	742	6,414	te [^]
13	1,183	6,204	no
14	6,021	14	ed [^]
15	5,955	920	ng
16	586	5,550	va
17	1,239	5,315	mo
18	25	5,288	no [^]
19	2,088	5,211	ss
20	5,161	440	ing
21	1,149	4,778	am
22	4,463	37	g [^]
23	4,442	1	y [^]
24	10	4,371	mo [^]
25	1,850	4,370	to
26	4,353	2	ng [^]
27	9	4,322	ti [^]
28	642	3,914	ere
29	1,002	3,509	sc
30	425	3,472	ass
31	961	3,452	vi
32	57	3,417	ava
33	3,374	101	n [^]
34	889	3,223	tt
35	493	3,162	ate [^]
36	1,250	3,112	ci
37	3,103	200	t [^]
38	3,071	5	es [^]
39	503	3,032	era
40	1,229	2,987	ai
41	210	2,874	[^] ri

42	880	2,749	Sa
43	3	2,710	ai^
44	780	2,649	Do
45	52	2,637	to^
46	42	2,612	ta^
47	2,574	784	Ion
48	2,573	225	Ea
49	276	2,561	Cc
50	2,545	853	r^

As to the English part, *n*-gram statistics were produced by using the combined *ICAME-BNC* word list. As to the Italian part, *n*-grams were generated from an Italian word list extracted from a random subset of 9.5 million word forms, further skimmed by considering only lower-case words, taken from *La Repubblica*.⁹

Only *n*-grams equal or longer than three characters and up to seven characters – including the apex symbol ‘^’ – were considered. In addition, only data displaying *n*-gram frequencies equal or higher than 80% of the total for one of the two languages (i.e. English and Italian) were used. Each word in the *HF* list was checked against this *n*-gram list and identified as a possible English or Italian word on the basis of the sum of the numbers when every *n*-gram sequence was checked (Table 2).

Table 2: Example of *n*-gram identification for the word *reason*.

<i>n</i> -gram sequence	<i>n</i> -gram frequency (English)	<i>n</i> -gram frequency (Italian)
on^	1,673	22
rea	587	104
eas	275	18
reas	72	10
total	2,607	154
percentage	94%	6%

Italian words were subsequently eliminated by subtracting the combined *OTACORAL* list used in the previous method. All words with accented characters or shorter than four letters were excluded from this provisional list of 3,863 items and a final list of 3,751 items was reached (Figure 6).

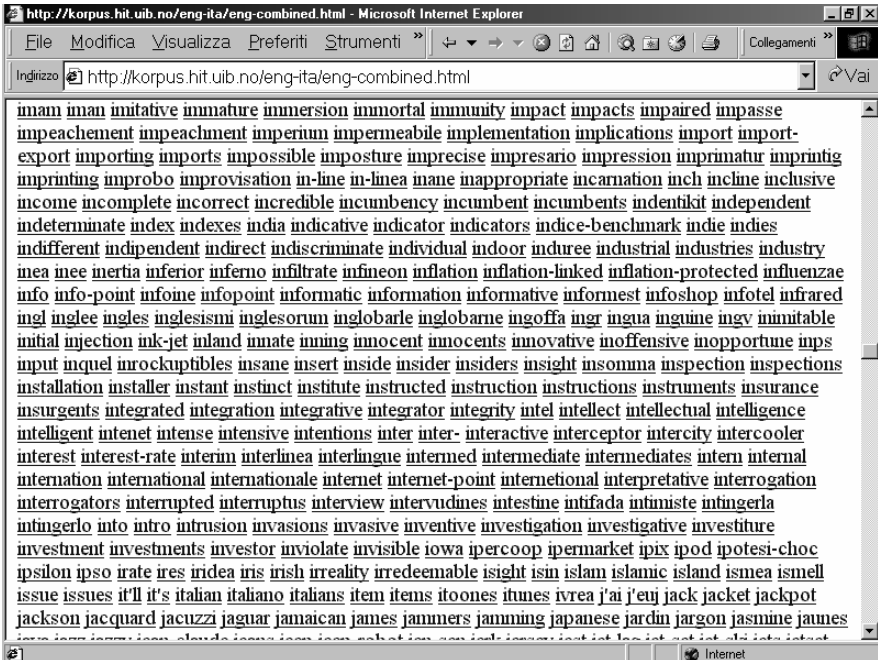


Figure 7: The searchable word list extracted from the *HF Corpus*.

When the user clicks on the underlined item, the software displays the concordances corresponding to the searched false anglicism. In the examples of *infopoint*, it is also possible to see the exact date on which the word appeared and the newspaper from which it was taken (Figure 3).

6. Some new false anglicisms found in the *HF Corpus*

The search for new false anglicisms was carried out by manually checking the context of occurrence of each single item belonging to the refined word list. This list proved to be useful mainly to find instances of autonomous compounds, i.e. words that do not have a formal equivalent in English, but semantic shifts and compound ellipses, i.e. words that have a formal equivalent in English, could not be readily recognised. Consequently, many English/Italian homographs had to be individually examined to verify whether they were authentic or false anglicisms.

After a preliminary analysis, this combined approach enabled the researchers to find some new false anglicisms, i.e. *barwoman*, *infopoint*, *stripman*, and *stripwoman*. Here follows a description of these false anglicisms with examples taken from the *HF Corpus* (see Figure 2). The false anglicism *barwoman* (e.g. ‘Studia Scienze delle Comunicazioni ma nel weekend fa la *barwoman*.’ LR040122), which in Italian refers to a woman who serves

beverages at a bar, corresponds to the English words *barmaid* or *bartender*. Italian speakers may not be aware of the fact that in English *maid* is used instead of *woman* to make the feminine of *barman*, whereas *bartender* is invariable for male and female.

Other false anglicisms retrieved are *stripman* and *stripwoman*. Similarly to *barwoman*, the appropriate substitute for *stripman* (e.g. ‘Con buffet, animazione intrigante e numerosi show e *stripman*.’ CS040128) and *stripwoman* (e.g. ‘È il piatto forte del locale: la table dance che ogni sera viene proposta agli ospiti da sexy *stripwoman*.’ CS031117) is simply *stripper*, since English does not differentiate between male and female strippers. Other appropriate English equivalents are *striptease artist* and *strip teaser*. It must be noticed that the plural *stripmen* was also found (e.g. ‘I partecipanti sono gli *stripmen* che vengono votati dal pubblico del locale.’ CS04029).

Finally, *infopoint* is the compound of *info(rmation)* – which is already a form of clipping – and *point* (e.g. ‘Ci sarà anche un *infopoint* per aggiornare i cittadini sul recupero dell’anfiteatro.’ CS030913). In this case, several English equivalents are available: *help desk*, *information booth*, *information bureau*, *information center*, and *information desk*.

7. Conclusion

The retrieval of false anglicisms from a corpus of Italian newspaper language cannot exclusively rely upon automatic processing. Manual scanning must be combined with computational procedures. As to the *HF Corpus*, the approach allowed the retrieval of up-to-date examples of false anglicisms of any type but it is only suitable to find new false anglicisms constituted by autonomous compounds written as single words.¹⁰ Although some automatic filters were added in order to eliminate the undesired ‘noise’ in the final word list, only further time-consuming manual scanning of such a list led to the tracing of new false anglicisms.

Undoubtedly, computational linguistic tools proved to be extremely useful to save time in building a corpus, in retrieving specific items, and in collecting a provisional list of false anglicisms in Italian. Despite the advantages, the computational techniques employed still do not seem to be sufficient to handle the complex and manifold phenomenon of false anglicisms. Along with automatic processing, a final manual scanning was indispensable.

Notes

- 1 False anglicisms encountered in press articles may be marked by single quotation marks (e.g. ‘recordman’), double inverted commas (e.g. “beauty farm”), double angle brackets (e.g. «food valley»), or italicised

orthography (e.g. *beauty case*). However, they are not always graphically signaled (e.g. footing).

- 2 The only cases in which graphic adaptation is attested comprehend false anglicisms which formally appear as compounds. In fact, there may usually be three alternative patterns: connected forms (e.g. *recordman*), compounds separated by space (e.g. *beauty case*), or hyphenated compounds (e.g. *block-notes*). At times, the same false anglicism may show the three alternatives simultaneously (e.g. *longseller*, *long seller*, and *long-seller*).
- 3 “It is clear that the words that one can most clearly label as ‘bogus’ are those that have no formal equivalent in English, and never have had [...]” (Spence 1987: 181).
- 4 According to the *GDU (Grande Dizionario Italiano dell’Uso)*, the autonomous compound *camera car* is used in motorcycle and car racing to indicate a special camera located on the vehicle while racing.
- 5 The implementation of the project described in this article has been accomplished thanks to the resources available at the *BATMULT (Bergen Advanced Training Site in Multilingual Tools)* located in the *Aksis* center (*Avdeling for kultur, språk og informasjonsteknologi*), formerly *HIT (Senter for humanistisk informasjonsteknologi)*, of the *University of Bergen*, Norway. The stay at *Aksis* was partly sponsored by the European Community through the *Marie Curie Training Site (MCTS)* host fellowship. The authors of the present article are solely responsible for the information published and do not represent the opinion of the Community. The Community is not responsible for any use that might be made of data appearing here.
- 6 The self-expansion of the *HF Corpus* mirrors the procedures developed for the *NNC (The Norwegian Newspaper Corpus)*. See Hofland (2000) for a detailed description of the automatic collection of newspaper articles from the Web and their inclusion in a corpus.
- 7 The lemmatised Italian verb list, which was extracted from a morphological lexicon project developed through the *CORAL* program at the *Università degli Studi di Bologna*, was made available by Guy Aston and Marco Baroni.
- 8 See Krenn and Samuelsson (1997) and Caropreso et al. (2001) for a comprehensive definition of *n*-grams and their use in Computational and Corpus Linguistics.

- 9 The Italian word list, from which *n*-gram statistics were produced, was extracted from a random subset of 9.5 million word forms taken from the newspaper *La Repubblica* and belonging to a text categorization project developed through the *CORAL (Corpora e apprendimento linguistico)* program at the *Università degli Studi di Bologna*. Thanks are due to Guy Aston and Marco Baroni for allowing access to this resource.
- 10 See Furiassi (2003) for a preliminary list of previously found false anglicisms.

References

- Bevitori, C. (2002), 'Le altre lingue e l'inglese: prestiti linguistici e risorse elettroniche', in: F. San Vicente (ed.) *L'inglese e le altre lingue europee. Studi sull'interferenza linguistica*. Bologna: CLUEB. 51-66.
- Caropreso, M. F., S. Matwin and F. Sebastiani (2001), 'A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization', in: A. G. Chin (ed.) *Text databases and document management: theory and practice*. Hershey (PA): Idea Group Publishing. 78-102.
- Furiassi, C. (2003), 'False anglicisms in Italian monolingual dictionaries: a case study of some electronic editions', *International journal of lexicography*, 16: 121-142.
- Hofland, K. (2000), 'A self-expanding corpus based on newspapers on the Web', in: *Proceedings of the second international language resources and evaluation conference (LREC)*, Athens, Greece, May 2000. Paris: European Language Resources Association (ELRA) [online] Available from: <http://gandalf.hit.uib.no/non/lrec2000/pdf/362.pdf>.
- Iamartino, G. (2001), 'La contrastività italiano-inglese in prospettiva storica', *Rassegna italiana di linguistica applicata*, 33: 7-130.
- Krenn, B. and C. Samuelsson (1997), *The linguist's guide to statistics*. Universität des Saarlandes: Computerlinguistik [online] Available from: http://www.coli.uni-sb.de/~krenn/stat_nlp.ps.gz.
- Marello, C. (1996), *Le parole dell'italiano: lessico e dizionari*. Bologna: Zanichelli.
- Pulcini, V. (2002), 'Italian', in: M. Görlach (ed.) *English in Europe*. Oxford: Oxford University Press. 151-167.
- Spence, N. C. W. (1987), 'Faux amis and faux anglicismes: problems of classification and definition', *Forum for modern language studies*, 23: 169-183.

Software

Corpus Workbench (CWB)

Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart
Arne Fitschen (ed.)
<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>

w3mir

Institutt for Informatikk, Universitetet i Oslo
Nicolai Langfeldt (ed.)
<http://langfeldt.net/w3mir/>

Corpora

British National Corpus (BNC)

Oxford University Computing Services
<http://www.natcorp.ox.ac.uk>

HF Corpus

Aksis, Universitetet i Bergen
Knut Hofland and Cristiano Furiassi (eds.)
<http://korpus.hit.uib.no/cwb/ita.html>

International Computer Archive of Modern and Medieval English (ICAME)

Aksis, Universitetet i Bergen
<http://www.hit.uib.no/icame.html>

Norwegian Newspaper Corpus (NNC)

Aksis, Universitetet i Bergen
Knut Hofland (ed.)
<http://avis.uib.no/>

Databases

Corpora e Apprendimento Linguistico (CORAL)

Scuola Superiore di Lingue Moderne per Interpreti e Traduttori (SSLMIT)
Università degli Studi di Bologna
<http://sslmit.unibo.it/~coral/>

Oxford Text Archive (OTA)

University of Oxford
<http://ota.ahds.ac.uk/>

Newspaper web sites

<i>Il Corriere della Sera (CS)</i>	www.corriere.it
<i>La Repubblica (LR)</i>	www.repubblica.it
<i>La Stampa (LS)</i>	www.lastampa.it

This page intentionally left blank

Lexical semantics for software requirements engineering – a corpus-based approach

Kerstin Lindmark*, Johan Natt och Dag**, Caroline Willners*

* Lund University

** Lund Institute of Technology

Abstract

In companies that constantly develop new software releases for large markets, there continually arrive new requirements, written in natural language that may affect the development work. Before any decision is made about the requirements, these must be analysed and understood, and related to the current set of implemented and queued requirements. This task is time-consuming owing to the high inflow of requirements, and decision-making would be facilitated by any support that would reduce the requirements analyst's workload. One of the main tasks is finding requirement duplicates and requirements with similar content and different NLP methods have been tried for this. Simple word matching is one of the methods used for linkage between requirements. If links could be set up not only between words, but also between concepts at different semantic levels, the chances of finding content-corresponding requirements would be greater. One goal of this project is to establish a terminology for requirements as well as to establish (Wordnet-type) semantic relations between terms, in order to enable multi-level linkage. For this purpose, we use a corpus consisting of 1,932 authentic software requirements, written in English of varying grammatical and stylistic quality. First, term candidates were extracted using the WordSmith Keyword function, with BNC Sampler as reference corpus. To find out whether there is any terminology specific to the 'requirements' sub-domain of the 'software' domain, the documentation associated with the software to which the requirements relate was also used as a reference (separately). Then, lexico-semantic patterns according to Hearst (1992) were used to find hyponymy-hyperonymy relations, and to confirm manually established relations. These analyses were performed on the text both 'as is' and, reducing noise somewhat, after POS-tagging by means of the Brill tagger (Brown Corpus tag-set). The results so far suggest that corpus-based methods are of importance to the management or requirements analyses.

1. Introduction

In software development, requirements are an important means for meeting the needs of the customer. The requirements may be identified and specified either before the specific software has been designed, forming a point of departure for the software engineer, or after the software has been tested or used for some time, thus forming a basis for correcting and improving the software and meeting the needs and complaints of the customer.

In traditional software development, also known as 'bespoke' software development, requirements are usually negotiated and agreed upon with a known

end user or customer, before development is pursued (Sawyer 2000). Thus, the developing company and the customer together identify and specify the requirements. Further, the customer may provide feedback during development to influence which particular requirements should be implemented.

In market-driven software development, there is very limited negotiation with end users and customers (Potts 1995; Carlshamre 2002). Rather, requirements are invented and specified in-house. This means that the developing company relies on employees providing appropriate requirements and the employees act as stakeholders for different parts of the organisation. Stakeholders are, for example, marketing, support and development departments that provide requirements of different kinds, such as bug reports, suggestions for new functions, suggestions for improving already existing functions, etc.

In order not to miss any good ideas, anyone within the organisation may submit a requirement, and requirements are collected continuously during development. The requirements are mostly written in natural language and stored in some kind of database. Analysing the requirements is a time-consuming and tiresome task, and the number of requirements thus easily grows faster than they can be dealt with. Therefore, requirements analysis and management often become bottlenecks in the development process.

Different attempts at analysing the requirements more or less automatically have been undertaken (Natt och Dag, Regnell, Carlshamre, Andersson and Karlsson 2002: 21-22). These automated techniques could support the analyst by suggesting duplicates and groupings based on the linguistic content of the requirements. To perform a deeper linguistic analysis of software requirements, an adequate lexicon is crucial; the lexicon must be domain-specific and also contain everyday vocabulary. A manually constructed lexicon is rather expensive to produce and therefore less appealing to the software development company. In particular, the market-driven organisation does not interpret requirements in the same way as in customer-specific development. In the latter, developers usually have very good domain expertise, whereas in the former they more often rely on consultants and brainstorming sessions (Lubars, Potts and Richter 1993). Therefore, it is even more troublesome to construct an appropriate domain-specific lexicon manually.

Under the assumption that it is possible to extract domain-specific terms from a collection of natural language texts in an automated, rapid manner, without pre-treatment, the following questions form the basis for the current investigation:

1. A) Is there a specific terminology for software requirements? B) Does the terminology of requirements coincide with the terminology of other documentation within the domain? C) Is the terminology consistent over time?
2. Would it be possible to structure such a domain-specific vocabulary into *Wordnet* parts that could be of relevance?

3. Can the structuring of *Wordnet* be automatised to some extent?
4. Will a *Wordnet*-type lexicon be advantageous for industrial analysis and grouping of software requirements?

Within the framework of a joint technology-linguistics project planned at the departments of telecommunications and linguistics at the Lund Institute of Technology and Lund University, respectively, a *Wordnet*-type lexicon (Fellbaum 1998; Vossen et al. 1997) will be used as a basis for linguistic analyses of requirements. This paper, which is an extension of (Natt och Dag and Lindmark 2002), deals with the construction of such a lexicon using a semi-automated, corpus-based approach, summarising the approaches adopted and results obtained within the framework of the current project.

2. Data sources

Three data sources have been analysed and compared, with respect to the vocabulary used:

The BNC Sampler

The *British National Corpus (BNC)* (Burnard 2000) is a 100 million word corpus of modern British English, both spoken and written, created in 1994 by a consortium of leading dictionary publishers and research institutions. From the full *BNC*, a two percent sample has been created and called the *BNC Sampler*. The *BNC Sampler* consists of one million words spoken, and one million words written. More information is available at <http://info.ox.ac.uk/bnc>.

Software documentation

The documentation (*Telelogic Tau 4.2 Documentation* 2001), a manual, is written in regular English and should also comprise domain specific words. The documentation used belongs to version 4.2 of the software application and is thus comparatively mature. In total, it comprises 1,069,478 words (tokens), of which 14,085 are unique (types).

Software requirements

From the software vendor we received a database comprising 1,932 requirements written in natural language. The majority of the requirements are written in English, irrespective of the authors' mother tongue. Thus, the quality varies and in some requirements there are also Swedish phrases. Example requirements, with all the attributes that the company uses, can be found in Table 1a and Table 1b. Due to the continuous elicitation, the requirements concern different developmental stages, such as elicitation, selection, and construction, which means that the requirements are analysed to various degrees. The requirements have been collected during five years of development, starting in 1996. In

Table 2 the number of requirements from each year is shown together with the number of words they comprise.

Table 1a: Example requirement submitted 1996.

<i>RqId</i>	<i>RQ96-270</i>
Date	
Summary	Storing multiple diagrams on one file
Why	It must be possible to store many diagrams on one file. SDT forces to have 1 diagram per file. It's like forcing a C programmer to have not more than one function per file... The problem becomes nasty when you work in larger projects, since adding a procedure changes the system file (.sdt) and you end up in a mess having to 'compare systems'.
Description	Allow the user to specify if a diagram should be appended to a file, rather than forcing him to store each diagram on a file of its own.
Dependency	4
Effort	4
Comment	This requirement has also been raised within the multiuser prestudy work, but no deeper penetration has been made. To see all implications of it we should have at least a one-day gathering with people from the Organizer, Editor and InfoServer area, maybe ITEX? Här behövs en mindre utredning, en 'konferensdag' med förberedelser och uppföljning. Deltagare behövs från editor- och organizergруппerna, backend behövs ej så länge vi har kvar PR-gränssnittet till dessa.
Reference	
Customer	All
Tool	Don't Know
Level	Slogan
Area	Editors
Submitter	X
Priority	3
Keywords	storage, diagrams, files, multi-user
Status	Classified

Table 1b: Example requirement submitted 1997. With reference to its content, this is a duplicate of the requirement in Table 1a.

<i>rqId</i>	<i>RQ97-059</i>
Date	Wed Apr 2 11:40:20 1997
Summary	A file should support storing multiple diagrams
Why	ObjectGeode has it. It's a powerful feature. It simplifies the daily work with SDT. Easier configuration management. Forcing one file for each procedure is silly.
Description	The SDT 'Data model' should support storing multiple diagram on one file.
Dependency	4
Effort	1-2
Comment	Prestudy needed
Reference	http://info/develop/anti_og_package.htm
Customer	All
Tool	SDT SDL Editor
Level	Slogan
Area	Ergonomy
Submitter	X
Priority	3: next release (3.3)
Keywords	diagrams files multiple
Status	Classified

Table 2: Number of requirements from the different years and the number of words they comprise.

<i>year</i>	<i>requirements</i>	<i>words</i>
1996	459	34,588
1997	714	60,944
1998	440	40,099
1999	239	20,029
2000	80	7,911
Total	1,932	163,571

3. Term extraction

The process of establishing a terminology for a given domain involves not only a meticulous semantic analysis of terms and establishment of definitions, but also extraction of terms from a relevant material (Suonuuti 2001). Traditionally, this is time-consuming and involves a lot of manual work.

Whereas a terminologist normally spends a lot of time reading the material and trying to identify what words are typical for the domain, we decided to adopt a more mechanical approach.

Term extraction has been the subject of many investigations, using methods of different levels of automatisation that are appropriate for different purposes (Thurmair 2003). Most systems use statistics, shallow parsing or alignment of bilingual resources, and most resources are POS-tagged corpora. Since our corpus was not tagged, and since we wanted to use existing tools, we selected a commercially available corpus linguistic analysis tool to find the words and phrases which could be considered domain specific terms, and also general language expressions that appeared to be used in a specific way, or were over-represented.

3.1 Method

With *Wordsmith Tools*, a simple frequency list from the requirements corpus was first produced. This list was compared with the frequency list from the *BNC Sampler*, using the keywords feature in *Wordsmith Tools*. The resulting list of 500 words, whose frequency in the requirements word list was significantly higher than in the *BNC Sampler* word list, was then scrutinised manually.

The same procedure was repeated with, in turn, the documentation vs. the *BNC Sampler* list, the requirements vs. the documentation, subsets of requirements (the requirements from a certain year were extracted as a subset) vs. the whole set of requirements, the subsets vs. each other, each subset vs. a set consisting of the other subsets and, for each of these comparisons, a comparison in the opposite direction, i.e. the documentation vs. the requirements and so forth. It is, of course, to be expected that quite a number of the terms used in the requirements are multiword units. To investigate this, bigrams, trigrams, and tetragrams were extracted using the *n*-gram finding feature in *Wordsmith Tools* ('clusters'). As a default, this function finds *n*-grams with a frequency of at least two. Since this turned out to give neat, manageable lists of apparently relevant words and phrases, we decided to use these. The lists of bi-, tri- and tetragrams were compared to each other with the keyword feature in a way corresponding to what was done with the single words.

One of the crucial questions for this investigation was whether the terminology of the requirements is consistent throughout the different sets, making it possible to 'predict the linguistic content' of the whole set of requirements based on an analysis of a subset, and, even more important, of new requirements based on an analysis of the old ones. To investigate this, the requirements were divided into subsets, according to year of origin. All possible different combinations of subsets were analysed. A complete overview of all the comparisons that were found relevant for the questions and conclusions in this paper is found in Table 3. The complete resulting data for the analysis can be found in an Excel workbook at <http://www.telecom.lth.se/Personal/johannod/education/cling/AnalysisResults.xls>.

Table 3: Overview of comparisons between the different corpora and subsets of the corpora.

<i>id</i>	<i>corpus under investigation</i>	<i>reference corpus</i>
req→bnc	All requirements	<i>BNC Sampler</i>
req→doc	All requirements	Documentation
Doc→bnc	Documentation	<i>BNC Sampler</i>
2req→2bnc	Bigrams All requirements	Bigrams <i>BNC Sampler</i>
4req→4doc	Tetragrams All requirements	Tetragrams Documentation
97→96	Requirements 1997	Requirements 1996
98→96-97	Requirements 1998	Requirements 1996-1997
99→96-98	Requirements 1999	Requirements 1996-1998
00→96-99	Requirements 2000	Requirements 1996-1999
96→Rest	Requirements 1996	All requirements but 1996
97→Rest	Requirements 1997	All requirements but 1997
98→Rest	Requirements 1998	All requirements but 1998
99→Rest	Requirements 1999	All requirements but 1999
00→Rest	Requirements 2000	All requirements but 2000

3.2 Results and analysis

The keywords list resulting from the comparison of the full set of requirements and the *BNC Sampler* already shows some interesting features of the linguistic contents of these texts (Table 4, left column). Among the most significant words are not only domain-specific terms, but words with a high frequency in general language, such as *should*, *be*, and *is* (i.e. these are not considered terms at all. The words are boldfaced in the table and those that are adjectives are also italicized). This is not very surprising, considering that the requirements are about features that are not working or features that the requirement stakeholder would want to have. The comparison between the requirements and the documentation points even clearer in the same direction, where *I*, *should*, *would*, etc., are overrepresented words in the requirements (Table 4, right column).

Table 4: Truncated keywords lists from single word comparison between all the requirements, on the one hand, and the *BNC Sampler* and the documentation, respectively, on the other hand.

<i>N</i>	<i>req</i> → <i>bnc</i>	<i>req</i> → <i>doc</i>
1	SDL	I
2	SDT	SHOULD
3	SHOULD	WOULD
4	SYMBOL	SDT
5	FILE	IT
6	MSC	WE
7	EDITOR	TO
8	ORGANIZER	TODAY
9	USER	OUR
10	DIAGRAM	HAVE
11	FILES	CUSTOMERS
12	SYMBOLS	ITEX
13	CODE	DOCUMENTATION
14	TEXT	SUPPORT
15	ITEX	LIKE
16	BE	USER
17	MENU	MINISYSTEM
18	DIAGRAMS	<i>NICE</i>
19	SIMULATOR	THINK
20	PAGE	DON'T
21	DIALOG	ABLE
22	TOOL	MAKE
23	POSSIBLE	VERY
24	TAU	<i>EASIER</i>
25	IS	<i>BETTER</i>

The overrepresentation of expressions common in general language is shown even more clearly in the tetragram comparison (Table 5).

Table 5: Truncated keywords list from tetragram comparison between all the requirements and the *BNC Sampler*.

<i>N</i>	<i>4req</i> → <i>4bnc</i>
1	SHOULD BE POSSIBLE TO
2	IT SHOULD BE POSSIBLE
3	TO BE ABLE TO
4	ALLOW THE USER TO
5	IT WOULD BE NICE
6	SHOULD BE ABLE TO
7	IN THE DRAWING AREA
8	THERE SHOULD BE A
9	IS NOT POSSIBLE TO
10	IN THE SDL EDITOR
11	IT IS NOT POSSIBLE

12	MAKE IT EASIER TO
13	WOULD BE NICE TO
14	DEFECT POSTPONED IN #
15	PM DEFECT POSTPONED IN
16	WANT TO BE ABLE
17	IN THE MSC EDITOR
18	THE USER WANTS TO
19	SHOULD BE EASY TO
20	THE USER HAS TO
21	TO MAKE IT EASIER
22	IT IS POSSIBLE TO
23	I WOULD LIKE TO
24	LIKE TO BE ABLE
25	END # #
26	BE POSSIBLE TO USE
27	BE NICE TO HAVE
28	WOULD LIKE TO HAVE
29	THE PROBLEM IS THAT
30	MAKE IT POSSIBLE TO

Comparison of the tetragrams from the requirements to those in the documentation can further establish that the differences are not due to the particular domain, but rather to the type of text. The results from the comparison are shown in Table 6.

Table 6: Truncated keywords list from tetragram comparison between all the requirements and the documentation.

<i>N</i>	<i>Areq</i> → <i>Adoc</i>
1	SHOULD BE POSSIBLE TO
2	IT SHOULD BE POSSIBLE
3	TO BE ABLE TO
4	IT WOULD BE NICE
5	I WOULD LIKE TO
6	SHOULD BE ABLE TO
7	ALLOW THE USER TO
8	WOULD BE NICE TO
9	THERE SHOULD BE A
10	DEFECT POSTPONED IN #
11	WOULD LIKE TO HAVE
12	PM DEFECT POSTPONED IN
13	MAKE IT EASIER TO
14	WANT TO BE ABLE
15	TO MAKE IT EASIER
16	LIKE TO BE ABLE
17	THE PROBLEM IS THAT
18	BE NICE TO HAVE
19	END # #
20	IT WOULD BE VERY

21	WOULD LIKE TO BE
22	ID WAS # #
23	PREVIOUS ID WAS #
24	IT SHOULD BE EASY
25	WOULD BE NICE IF
26	I WANT TO BE
27	SHOULD BE EASY TO
28	USER SHOULD BE ABLE
29	THE USER WANTS TO
30	BE POSSIBLE TO USE

As for the actual domain-specific terms, these appear rather from the mono- and bigram lists (and, although to a lesser extent, also from the trigram lists, not shown here). In Table 7, the result from the comparison of single words lists of the requirements and the documentation, respectively, with the *BNC Sampler* is shown.

Table 7: Truncated keywords lists from single word comparison between all the requirements and the documentation, respectively, on the one hand, and the *BNC Sampler*, on the other hand.

<i>N</i>	<i>req</i> → <i>bnc</i>	<i>doc</i> → <i>bnc</i>
1	SDL	SDL
2	SDT	TELELOGIC
3	SHOULD	TAU
4	SYMBOL	PAGE
5	FILE	TYPE
6	MSC	FILE
7	EDITOR	USER'S
8	ORGANIZER	TTCN
9	USER	C
10	DIAGRAM	MANUAL
11	FILES	MARCH
12	SYMBOLS	UM
13	CODE	CHAPTER
14	TEXT	SUITE
15	ITEX	TEST
16	BE	SIGNAL
17	MENU	SYMBOL
18	DIAGRAMS	IS
19	SIMULATOR	DIAGRAM
20	PAGE	MENU
21	DIALOG	SYSTEM
22	TOOL	PROCESS
23	POSSIBLE	CODE
24	TAU	NAME
25	IS	COMMAND

Through the comparison of bigrams, domain-specific compound terms may be captured, although there are many non-useful combinations, as shown in Table 8.

Table 8: Truncated keywords lists from bigram comparison between all the requirements and the *BNC Sampler*.

N	2req→2bnc
1	SHOULD BE
2	THE USER
3	THE ORGANIZER
4	POSSIBLE TO
5	THE SDL
6	THE TEXT
7	IT SHOULD
8	BE POSSIBLE
9	THE MSC
10	THE SIMULATOR
11	IN THE
12	TO USE
13	BE ABLE
14	MACRO MINISYSTEM#
15	AN SDL
16	IS NOT
17	THE SYMBOL
18	SDL SYSTEM
19	THE EDITOR
20	ABLE TO
21	THE FILE
22	THE SDT
23	IN SDT
24	IN SDL
25	POSSIBILITY TO
26	USER TO
27	SDL EDITOR
28	CODE GENERATOR
29	THE TOOL
30	SDT #

The question whether there is a specific language for requirements is thereby accounted for. The keywords list compared to the *BNC Sampler* comprises many domain-specific words (Table 4). Although this is not surprising, the result validates what was expected at the outset. Further, the tetragram comparison (between requirements and the *BNC Sampler*) shows that certain phrases are used more often in requirements than in general language (Table 5). Actually, these phrases are not due to the domain but, rather, to the nature of requirements (Table 6).

There is an overlap between the documentation and the requirements, with respect to domain-specific terms. There is a difference, though, with respect to

other linguistic features. A comparison of both the keywords lists of the requirements and of the documentation to the *BNC Sampler* shows that the same domain-specific terms occur in both the requirements and the documentation (Table 7). The tetragram comparison between the requirements and the documentation shows, as already discussed above, that the type of phrases are due to the nature of requirements. In other words, requirements differ in the use of phrases (Table 6).

As to the question of the consistency of requirements throughout the different sets, some relevant results are shown in Table 9 and Table 10. Table 9 lists the terms that are significantly more and less common, respectively, in the requirements from 1996 compared to those from 1997. Table 10 shows the corresponding results from the comparison of the requirements from 1996 to those from all the other years. In brief, comparisons of one subset to the rest of the requirements yield a small number of keywords; in the case of bi-, tri-, and tetragrams there were none in quite a few cases. The small number can be even further reduced, if only words that could be term candidates are retained. As for the comparison of one subset with a set consisting of all the other sets but the one under investigation, this yields a somewhat larger number of keywords; approximately thirty per set.

It can be seen that some terms are unique for the 1996 set (terms marked with an asterisk). However, most of them may be disregarded as domain-specific terms. It may be concluded from these lists that the words used in requirements are quite consistent over the years.

A very interesting result is that the lists are relatively short, indicating that the differences between the words used in the requirements from 1996 and the words used during the subsequent years are limited.

Table 9: Complete list of keywords that are significantly more common in 1997 than 1996 and vice versa.

<i>More common 1997 than 1996</i>	<i>More common 1996 than 1997</i>
SYMBOL	ENV
CMICRO	ITEX
PAGE	SEND
HMSC*	SPEC
TEXT	N
SHOULD	
FLOW	
MSC	

*No occurrence 1996.

Table 10: Complete list of keywords that are significantly more common in 1996 than in the other years and a list of keywords that are significantly more common in 1997-2000.

<i>More common 1996 than other years</i>	<i>More common other years than 1996</i>
MINISYSTEM*	TELELOGIC
MACRO	H
N	FILES
LNO*	TEXT
AB	PAGE
SPEC	TAU
CALLER*	SYMBOL
SDL	
SEND	
SHALL	
ANSWER	
INMPOVEMENT*	
ACTIVEX*	
ROOT	
SDT	
TOOLBARS*	
RECEIVE	
AR*	
THANK*	
TIMER	
AWARE	
GATES	
COMMAND	
DEPENDENCY	
OOA	
SIMUI	

*No occurrence other years

4. Construction of the domain-specific *Wordnet*

The domain-specific vocabulary being established to some extent, the actual *Wordnet* was to be constructed. *Wordnet* construction has been an important issue for the authors within the Swedish *Wordnet* project (Viberg et al. 2002), but in this paper, only the finding of semantic relations between words or phrases in a semi-automatic manner is discussed.

4.1 Method and results: hyponymy and hyperonymy

Automatic, or semi-automatic, identification of semantic relations between words in a corpus has been the subject of a not negligible amount of work. Already in 1970, Robison (cited in e.g. Morin 1998) presented a model with lexical patterns indicating hyponymy relations. Hearst (1992) is a most important work in this

field. He indicates five lexical patterns in English that are typical for the hyperonymy relation:

- NP such as NP1(, NP2, and NP3)
- NP (, NP...NP) and other NP
- NP (, NP...NP) or other NP
- NP, especially NP
- NP, including NP (, NP and NP)

For the first pattern, this yields the relations hyperonym (NP, NP1), hyperonym (NP, NP2) and hyperonym (NP, NP3).

In this investigation, Hearst's model was applied and yielded surprisingly good results, considering the circumstances. In most, if not all, related works, the material consists of a part-of-speech-tagged corpus. In the current case, the corpus was not tagged at all (to begin with). The main reason for not tagging the corpus initially was that we wanted to see what could be done with the text as such. The requirements corpus was tagged later, though, for comparison.

The principle according to Hearst consists of extracting from the corpus certain lexical patterns, or syntactic constructions, indicating a semantic relation. The patterns shown in Hearst (1992) have been investigated (using *WordSmith Tools*). In the first stage, without POS-tagging, the combinations of words were searched as concordances, and their contexts were considered. For the expressions *and other*, *or other* and *such as*, good results were obtained, while *especially* and *including* yielded a lot of noise. This is hardly surprising, since the first three patterns can be used in conjunction with noun phrases only (practically), while *especially* and *including* can occur in many different contexts – and really do.

Table 11: Number of matches for the different patterns, and number of hyponyms extracted by help of the matches.

<i>Pattern</i>	<i>Requirements</i>	<i>Hyponyms</i>	<i>Documentation</i>	<i>Hyponyms</i>
And other	18	22	53	22
Especially	47	5	53	5
Including	27	6	245	6
Or other	4	5	11	5
Such as	11	41	144	41

The result seems surprisingly good, considering the small amount of preparatory work. However, a great deal of cleaning-up had to be done, and the question whether this work would have been considerably easier, had the corpus been POS-tagged, was raised. Eventually, it seemed reasonable to try this out in practice. Thus, the requirements corpus was tagged with the *Brill* tagger (<http://www.cs.jhu.edu/~brill/>), using the *Brown Corpus* tag set (<http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html>). Preparatory work

included tokenisation, so that punctuation marks and the like were surrounded by spaces, as well as breaking lines before every new sentence.

In Table 12, the results of the investigations of the tagged and the non-tagged version, respectively, are shown side by side.

Table 12: Hyponyms and hyperonyms obtained by searching for lexico-syntactic patterns before and after POS-tagging the corpus. Columns 4 and 5 show the results from the non-tagged version. The last lines show hypo- and hyperonyms that occurred only in the non-tagged version.

<i>Results – tagged text</i>			<i>Results – non-tagged text</i>	
<i>Hyponym</i>	<i>Hyperonym</i>	<i>Pattern</i>	<i>Hyponym</i>	<i>Hyperonym</i>
TTCN	languages/ ¹ notations	and other	TTCN	languages/ notations
MSCs	diagram types		MSCs	diagram types
Tau	Windows applications		Tau	Windows applications
System	SDL entities		System	SDL entities
Block	SDL entities		Block	SDL entities
Process	SDL entities		Process	SDL entities
MSC Editor	Editors		MSC Editor	other editors
SDT/ITEX	applications		SDT/ITEX	applications
Search	operations		Search	operations
Update Headings	operations		Update Headings	operations
Framemaker	tools		framemaker	tools
TTCN	documents		TTCN	documents
SDL	documents		SDL	documents
quick buttons	graphical icons	, and other		
Document icons	graphical icons		document icons	graphical icons
special buttons	graphical icons		special buttons	graphical icons
tutorials	descriptive material		tutorials	descriptive material
methodology guidelines	descriptive material		methodology guidelines	descriptive material
organizer log	tools		organizer log	tools
text editor	tools		text editor	tools
SDT	Applications	and/or other	SDT	applications
Word	tools	or other	Word	tools

C	textual languages	such as	C	textual languages
C++			C++	textual languages
basename			basename	commands
Alt			Alt	In-line expressions
Opt			Opt	In-line expressions
Exc			Exc	In-line expressions
Loop			Loop	In-line expressions
UnsignedChar			UnsignedChar	types
Unsigned LongInt			UnsignedLongInt	types
UnsignedLon			U????	types
tab			tab	shortcut
Erisoft			Erisoft	customers
Hagenuk			Hagenuk	customers
XMP			XMP	XMP
TCP/IP			TCP/IP	TCP/IP
Sockets			Sockets	Sockets
XOM	encoding decoding algorithms		XOM	encoding decoding algorithms
XDR	encoding decoding algorithms		XDR	encoding decoding algorithms
BER	encoding decoding algorithms		BER	encoding decoding algorithms
PER	encoding decoding algorithms		PER	encoding decoding algorithms
DER	encoding decoding algorithms		DER	encoding decoding algorithms
Borland	PC IDEs		Borland	PC IDEs
Microsoft	PC IDEs		Microsoft	PC IDEs
Copy	menu choices		Copy	menu choices
Insert	menu choices		Insert	menu choices
Replace	menu choices		Replace	menu choices
Alta/BoNES	tool vendors		Alta/BoNES	vendors

OPNET	tool vendors		OPNET	vendors
ClearCase	software configuration management systems			
DOORS	development tools	, such as	DOORS	development tools
ClearCase	development tools		ClearCase	development tools
Visigenics	ORBs			
Chorus	ORBs			
buttons	components	(such as	buttons	components
text fields	components		text fields	components
task symbol	symbol			
projector	display		projector	display
Comic Sans	font	, such as	Comic Sans	font
C files	generated files		C files	generated files
setup	secondary dialog	(such as	setup	dialog
the INCREMENT lines	lines of the license	Including		
submodules	modules			
state symbol	symbols	, including	state symbol	symbols
bug fixes	changes	(including	bug fixes	changes
GEODE	tools			
hyperlinks	List notes			
ITEX	Telelogic tools/NNS	, including	ITEX	Telelogic tools
			SDT windows	windows
			SDL	diagram
			MSC	diagram
			OOA	diagram
		such as	ASN.1	extensions
			*.tsp	files
			*.pr	files
			*.err	files
		especially	ctypes.sdl	predefined sorts in SDT
			Win32	OS'es
			Windows NT	OS:s
			CORBA	OS:s
			Win32	OS'es
			Graphical	names

			CORBA Tracer	
			print	commands

The table shows that the tagged version yielded eight relations that were not captured in the non-tagged version, and that the version without tags yielded 15 relations that were not captured in the tagged version. Tagging did not improve the result very much, although some cleaning-up work was avoided.

4.2 Meronymy

Within the domain of software engineering, the meronymy relation seems to be just as frequent as the hyponymy relation. For instance, many concepts refer to different modules and components of a software system, parts of the screen and subsets of programming languages. The manual term grouping (not presented here) also showed that this is the case, but no meronymy relations have been uncovered by way of lexico-syntactic patterns. In the literature, practically only hyponymy-hyperonymy relations are mentioned in connection with pattern-matching, so it can be assumed that other semantic relations are hard to find this way.

In this study, some searches were performed in vain. For example, all the patterns in table 13 were tried out on the requirements corpus. The results were very discouraging and are not presented here.

Table 13: Examples of patterns investigated for meronymy.

BELONG TO	INCLUDED
BELONGING TO	INCLUDES
BELONGS TO	INCLUDING
BETWEEN	INTERNAL
CONSISTS OF	IS INCLUDED
CONTAINED IN	IS PART
CONTAINING	LOCATED
CONTAINS	MADE UP
FOUND IN	MAKE UP
HAS A	OF THE
HAS AN	PLACED
HAVING	WHOLE
HAVING	WITHIN
IN	

Both explicit and implicit meronymies were searched for without success. Some examples have been found, but none that could be used. The limited size of the material could be an explanation, but since other works present failed investigations as well, this is probably not the case.

To conclude, patterns indicating hyponymy relations can be said to yield useful results even in a non-tagged corpus, i.e. even without syntactic specifications and limitations, while patterns indicating meronymy relations have not really been found.

5. Conclusions

The assumption that it is possible to extract domain-specific terms from a collection of natural language texts automatically proved to be correct. By using the KeyWords function in *Wordsmith Tools*, words and phrases characterising a domain-specific corpus can be found surprisingly rapidly. There are some reservations, though: The amount of noise is rather substantial, resulting in extensive post-editing. Preparatory work including lemmatising and part-of-speech-tagging could be expected to improve quality and speed.

To conclude, we return to the questions asked initially:

1. A) Is there a specific terminology for software requirements? B) Does the terminology of requirements coincide with the terminology of other documentation within the domain? C) Is the terminology consistent over time?

Yes. Firstly, the comparison with the *BNC Sampler* yielded a great number of domain-specific expressions. Secondly, the tetragram comparison (requirements and *BNC Sampler*) shows that certain phrases are more frequent in requirements than in general language. Thirdly, the occurrences of these phrases are not related to the domain as such, but to the type of text represented by the requirements. The terminology is consistent between requirements and documentation, but the text type makes a difference also between these two text collections, which is reflected in the set of most frequent phrases, especially in the tetragrams. As for the consistency over time, Table 9 shows that the terms and phrases are the same to a very large extent.

2. Would it be possible to structure such a domain-specific vocabulary into *Wordnet* parts that could be of relevance?

Yes, or at least, it is possible to establish smaller sub-nets for different fields within the domain. The adding of other terms will be necessary for the nets not to be too fragmentary.

3. Can the structuring of *Wordnet* be automatised to some extent?

Automatic extraction of relations by way of lexico-syntactic patterns has proved helpful for establishing semantic relations between concepts. This is true especially for hyponymy–hyperonymy. In this case, the pattern matching is a

rapid and efficient approach, not requiring a great deal of preparatory work. Part-of-speech-tagging turns out not to be necessary. Unfortunately, meronymy relations do not seem to be easily captured in this way.

4. Will a *Wordnet*-type lexicon be advantageous for industrial analysis and grouping of software requirements?

Since we have not performed any practical tests, this question can hardly be answered yet. However, a *Wordnet*-type lexicon could be assumed to assist in identifying requirements with similar content, even if they differ on the surface-level. Searching with matching of hyponymy-hyperonymy-related concepts might not directly yield duplicates as a search result, but requirements related to the same phenomenon should be captured. (This is valid even more for meronymy-related concepts.) If, for example, a requirement concerns 'storing diagrams' and another requirement concerns 'storing MSC charts', then an automatic analysis based on word-matching only would not be able to link these requirements, whereas a system recognising hyponyms ('MSC chart') and hyperonyms ('diagram') would signal that these requirements have something in common, and the requirements analyst could start to compare them at once, instead of having to read a large number of requirements and react to the similarity of content him- or herself. If this works, then the *Wordnet*-type lexicon has contributed to the time-savings aimed for, and the purpose of the current work has been fulfilled.

Note

- 1 The tagger tagged the slash as a noun: "/NN".

References

- Burnard, L. (ed.) (2000), *The British National Corpus users reference guide*. Oxford: Oxford University Computing Services, British National Corpus.
- Carlshamre, P. (2002), *A usability perspective on requirements engineering – from methodology to product development* (Dissertation No. 726). Linköping: Linköping University, Linköping Studies in Science and Technology.
- Fellbaum, C. (ed.) (1998), *Wordnet – an electronic lexical database*. Cambridge (MA): MIT Press.
- Hearst, M. (1992), 'Automatic acquisition of hyponyms from large text corpora', in: *Proceedings of the 14th international conference on computational linguistics*. Nantes, France, 20-28 July 1992. 539-545.
- Lubars, M., C. Potts and C. Richter (1993), 'A review of the state of the practice in requirements modelling', in: *Proceedings of IEEE international symposium on requirements engineering*, San Diego (CA), USA, 1-6 January 1993. Los Alamitos (CA): IEEE Computer Society Press. 2-14.

- Morin, E. (1998), 'Prométhée: un outil d'aide à l'acquisition de relations sémantiques entre termes', in: P. Zweigenbaum (ed.) *Actes de la 5e conférence annuelle sur le traitement automatique des langues naturelles*, Paris, France, 10-12 juin 1998. 172-181.
- Natt och Dag, J. and K. Lindmark (2002), *Selecting an appropriate language base for automated requirements analysis* [online]. Available from: http://www.cs.lth.se/%7Eepierre/CLCourse/Reports/johan_kerstin.pdf.
- Natt och Dag, J., B. Regnell, P. Carlshamre, M. Andersson and J. Karlsson (2002), 'A feasibility study of automated support for similarity analysis of natural language requirements in market-driven development', *Requirements engineering journal*, 7 (1): 20-33.
- Potts, C. (1995), 'Invented requirements and imagined customers: requirements engineering for off-the-shelf software', in: *Proceedings of the second IEEE international symposium on requirements engineering*, York, England, 27-29 March 1995. Los Alamitos (CA): IEEE Computer Society Press. 128-130.
- Robison, H. R. (1970), 'Computer-detectable semantic structure', *Information storage and retrieval*, 6: 273-288.
- Sawyer, P. (2000), 'Packaged software: challenges for RE', in: *Proceedings of sixth international workshop on requirements engineering: foundation for software quality*, Stockholm, Sweden, June 2000. Essen: Essener Informatik Beiträge. 137-142.
- Suonuuti, H. (2001), *Guide to terminology* (2nd ed.). Helsinki: Tekniikan sanastokeskus.
- Thurmair, G. (2003), 'Making term extractions tools usable', in: *Controlled language translation, EAMT-CLAW-03*. Dublin City University, Dublin, Ireland, 15-17 May 2003 [online]. Available from: <http://www.eamt.org/archive/dublin/THURMAIR.PDF>.
- Viberg, Å., K. Lindmark, A. Lindvall and I. Mellenius (2002), 'The Swedish Wordnet project', in: A. Braasch and C. Povlsen (eds.) *Proceedings of the tenth EURALEX international congress, vol. II*, Copenhagen, Denmark, 13-17 August 2002. Copenhagen: Center for Sprogteknologi. 407-412.
- Vossen, P., L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge and W. Peters (1997), *The EuroWordnet base concepts and top ontology*. EuroWordnet (LE 4003) Deliverable D017, D034, D036. University of Amsterdam.