Typological Studies in Language 82

Formulaic Language

VOLUME 1

Distribution and historical change

edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen M. Wheatley

John Benjamins Publishing Company

Formulaic Language

Typological Studies in Language (TSL)

A companion series to the journal *Studies in Language*. Volumes in this series are functionally and typologically oriented, covering specific topics in language by collecting together data from a wide variety of languages and language typologies.

Volume 82

Formulaic Language. Volume 1. Distribution and historical change.
Edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali
and Kathleen M. Wheatley

Formulaic Language

VOLUME 1

Distribution and historical change

Edited by

Roberta Corrigan

Edith A. Moravcsik

Hamid Ouali

Kathleen M. Wheatley

University of Wisconsin-Milwaukee

John Benjamins Publishing Company

Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Formulaic language : volume 1 : distribution and historical change/ edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen M. Wheatley.
p. cm. (Typological Studies in Language, ISSN 0167-7373 ; v. 82)
Includes bibliographical references and index.
1. Linguistic analysis (Linguistics) 2. Linguistic models. I. Corrigan, Roberta.
P126.F67 2009
410--dc22 2008042109
ISBN 978 90 272 2995 3 (HB; alk. paper) – ISBN 978 90 272 2997 7 (SET: alk. paper)
ISBN 978 90 272 9017 5 (EB)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

This book is dedicated to the memory of our dear friend and colleague Michael Noonan (September 14 1947 – February 23 2009). The conference on formulaic language that these papers come from and the publication of this book would not have happened without Mickey's enthusiastic leadership, encouragement, and loving attention.

Table of contents

Preface

This two-volume collection presents revised versions of a selection of papers from the 25th UWM Linguistics Symposium on Formulaic Language, held on April 18–21, 2007 at the University of Wisconsin-Milwaukee. To our knowledge, it was one of the first conferences specifically devoted to this topic.

We are grateful to Joan Bybee, who suggested the topic for this conference, and to Michael Noonan, who took primary responsibility for organizing it. We gratefully acknowledge the funds provided by various units of the University of Wisconsin-Milwaukee – the Department of English, the Department of Foreign Languages and Linguistics, the Center for International Education, and the College of Letters and Science – as well as those that came from royalties derived from the Benjamins' book series "Typological Studies in Language" due to the generosity of the editors of the previous volumes of this series and of Cornelis Vaes of John Benjamins. Heart-felt thanks also to our colleagues, students, and office staff for their invaluable help in putting on this event.

The indices were prepared by Deborah Mulvaney. We are grateful to her for her work performed under difficult conditions.

This preface and the introductory paper to follow are included in both volumes.

Introduction. Approaches to the study of formulae

Roberta Corrigan, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley

1. What are formulae?

Languages generally afford their speakers considerable freedom in how to express their ideas. This freedom is twofold, extending both to the choice of elements and to their arrangement. Consider the examples in (1).

(1) a. *Bill fixed the faucet.*
b. *Bill repaired the faucet.*
c. *Bill repaired the spigot.*
d. *My brother fixed the faucet.*

Under appropriate conditions, all four sentences in (1) can express the same meaning. If fixing the faucet involved actually repairing it, (b) serves as a paraphrase of (a). If the speakers are familiar with both words *faucet* and *spigot*, (c) is a paraphrase of (a) and (b). And if Bill happens to be the speaker's brother, (d) is also a possible way of conveying the same meaning.

The sentences of (1) show that there are alternative **lexical items** for expressing the same meaning. The same holds for how sentences are **structured**.

(2) a. *Bill fixed the faucet last night.*
b. *Last night, Bill fixed the faucet.*
c. *The faucet was fixed by Bill last night.*
d. *It was Bill who fixed the faucet last night.*
e. *What Bill fixed last night was the faucet.*

What the sentences of (2) show is that there are also alternative grammatical structures that can be used to express a meaning. The choice among them is context-dependent but, in terms of truth value, the five are equivalent.

The freedom to choose forms for expressing something does not hold to the same extent on all levels of language structure. The examples of (1) and (2) illustrate the considerable freedom we have in constructing **sentences**.

On the one hand, the range of choices is much larger on the **discourse** level: the same event, for example, may be described by a different selection and sequencing of sentences. On the other hand, the range of allowable alternatives narrows as we proceed to the selection and arrangement of linguistic units smaller than the word. In constructing words, one **morpheme** generally cannot be replaced by another, even if both have a similar or identical meaning, nor can morpheme order be changed. (3) shows this for compounds, (4) shows it for derived words.

(3) a. *lighthouse*
b. **lightbuilding*
c. **houselight*

(4) a. *unpleasant-ness*
b. **unpleasant-icity*
c. **ness-unpleasant*

The fact that components of a word can generally not be replaced by other equivalent parts and that the order of the parts cannot be reversed is also true for meaningless **phonetic segments**. (5) illustrates that phonemes cannot be replaced by others nor can their order be changed with the meaning remaining the same, even if the variants are within the bounds of phonotactic constraints.

(5) a. *block*
b. **plock*
c. **cklob*

So far it would seem that, whereas in constructing words out of phonetic segments and out of morphemes, form variation is restricted or altogether absent, constructing sentences out of words and discourses out of sentences allows for a broad range of options. Sinclair (1991: 109) coined the phrase "the open choice principle" to describe the notion that text – sentences and discourses – can result from a large number of complex choices.

However, Sinclair (1991: 110) also called attention to the fact that certain kinds of text afford less freedom of choice. He contrasted the open choice principle with the "idiom principle", which states that texts generally include "a large number of semi-preconstructed phrases that constitute single choices, even though they may

appear to be analyzable into segments." (On the idiom principle, see also **Bybee and Cacoullos** (volume 1), and **Ellis and Frey, Erman,** and **Van Lancker Sidtis** (both in volume 2).) For example, consider (6).

(6) a. *This is water under the bridge.*
b. *He is pushing the envelope.*
c. *Try to think outside the box.*
d. *Mary spilled the beans.*

These expressions, just as those in (1), do allow lexical and structural alterations, but only if they are meant in their literal sense. As shown in (7), the altered versions have lost their metaphoric, idiomatic interpretation.

(7) a. *This is water below the bridge.*
b. *He is giving a push to the envelope.*
c. *Try to think outside the crate.*
d. *Mary spilled the garbanzo beans.*

These examples suggest that words of a sentence can be replaced and re-arranged as long as the sentence is compositional but in their idiomatic reading, this freedom is lost (Nunberg, Sag and Wasow 1994).

Is it generally true that compositionality is a necessary condition for alterable word choice and word arrangement? Consider (8).

(8) a. *The check is in the mail.* (response to an inquiry)
b. *Your call is important to us.* (voice mail message when the caller is put on hold)
c. *How can I help you?* (in a store)
d. *Are you OK?* (after a fall)
e. *I hear you.* (in a discussion)

These sentences are not idioms: they are compositional and, as shown in (9), they may be constructed in alternative ways.

(9) a. *We have placed the check in the mail.*
b. *Your telephone call has great importance to us.*
c. *How may I assist you?*
d. *I wonder if you have hurt yourself.*
e. *I understand what you are saying.*

The alternatives in (9) are all possible expressions but in the contexts indicated in (8), they are much less likely to be actually used. The respective meanings **could be** expressed differently from (8) but in fact they generally **are not**. In these cases, the speaker appears to renounce the great freedom that the language offers for alternative expressions of the same meaning and opts for a single format.

The expressions in (8) are prototypical examples of **formulae**. Two distinctive characteristics differentiate them from ordinary sentences: restricted form

and restricted distribution. **Restricted form** means formulae are not amenable to lexical and structural re-formulations. They are couched in only one of the several alternative ways permitted by the language, and only a single item – or a limited set of lexical items – can fill the structural slots. They are structurally rigid: they underutilize the resources made available by the language for expressing a particular meaning. In this respect, formulae are more like words and morphemes than ordinary sentences. From the point of view of relative rigidity of form, formulae and idioms form a single class. Idioms are a particular subclass within this broader category, characterized by non-compositionality.

Restricted distribution means formulae tend to occur in particular styles of language tied to particular communicative situations. Ordinary sentences may also be subject to stylistic constraints: what we can say and the words and structures that we use depend to an extent on the context. For formulae, however, meaning and form are jointly favored or disfavored in given situations. Thus, formulae may serve as true hallmarks of style. For example, the redundant phrase in *Chicago will be our last and final stop* evokes the voice of a public address system in planes or trains.

However, it should also be noted that restrictions on the form and the distribution of formulae are merely probabilistic rather than absolute. Formulae do tolerate some form variation and, while they may be favored in given contexts, they are not uniquely keyed to situations. As Wray (2002: 25) puts it, formulae are "preferred choices" for expressing certain meanings.

The formulae discussed in the papers of this book actually vary in how closely they conform to the prototype described above. At one end of the scale of structural rigidity are compounds, such as *lighthouse*, where both lexical material and linear order are fully fixed. At the other end are syntactic constructions such as topic phrases in Japanese (analyzed by **Kurumada** (volume 2)), where the only recurrent lexical item is the topic marker *wa*, with the following noun phrase freely chosen. **Ellis and Frey** (volume 2) present data on another type of formula that is at the less rigid end of the continuum. In semantic prosody, there is huge flexibility in what can combine with a target word, but more rigidity in whether the collocate is negative or positive in its affective evaluation. For example, *achieve* has positive prosody because it is most likely to occur with positive collocates such as *success* or *goals*. An example of negative semantic prosody is described by Corrigan (2004) who found that in conversations between parents and their young children, utterances surrounding the phrase *what happened?* were more likely to be negative than positive.

A range of structural rigidity can also be seen within constructions involving the same word. **Hudson and Wiktorsson** (volume 1) investigate the formulaic patterns of the relater *about* and argue that around 80% of the ADJ+*about* and NOUN+*about* datasets they studied can be described in terms of constructions – from the more substantive and highly idiomatic expressions (*thing about* X *is*, *sorry about that*),

which pattern to a large extent with meanings with a negative or generally unfavourable orientation, to the more schematic ([N] + *about*) where the noun belongs to one of a few sets (general noun, noun of mental state or activity, noun of opinion or communicating opinion).

In sum, we have described prototypical formulae as constructions that have restricted forms and restricted distributions. The papers in this book range widely in how closely they adhere to the prototype. (For alternative definitions of formulaicity and their applicability, see **Wray**'s paper in volume 1).

2. Research questions

The study of formulae is a timely endeavor: it fills a gap in today's linguistic research for two reasons.

First, grammatical work in the past few decades paid primary attention to the creative aspects of language. It has of course been recognized that, as in all other aspects of human creativity, the production of sentences, too, is subject to constraints: some things are allowable and others are not. But these constraints were researched on the highest, most general level. Less attention seems to have been paid, on the one hand, to utterances that stretch the limits of these constraints, such as individual idiosyncrasies or poetic language, and, on the other hand, to utterances that underutilize the freedom afforded by general constraints of the language, such as set phrases: formulae. Formulae represent the flip side of creativity in language: they utilize a narrowly defined set of choices from among all the alternatives that rules of discourse, syntax, morphology and the lexicon would allow for.
In sharp contrast to the creative aspects of the linguistic behavior of language-users, formulae attest to the imitative aspects of this behavior.

The frequency with which formulae occur has not been the focus of most work in generative grammar. Yet, in recent years, several studies have suggested that formulaic expressions are far more frequent than previous work had acknowledged. Cameron-Faulkner, Lieven, and Tomasello (2003) looked at the distribution of item-based phrases in English-speaking mothers' language directed to their children. Fifty-one percent of all the maternal utterances began with one of 52 item-based phrases. Erman and Warren (2000) found that 55% of spoken and written text is constructed out of formulae. In volume 2 of this book, **Bannard and Lieven** examine recurring strings of speech that two-year-old English-speaking toddlers have either used or heard previously. They find that only about 3 to 14% of the utterances could not be derived from previous strings.

The other reason why formulae have not been extensively studied is that, as noted in the preceding section, their structural and lexical characteristics elude absolute,

binary characterization. The choice of words and choice of structure in formulae do leave some latitude: they can be described only probabilistically. Similarly, rules about the stylistic and situational distribution of formulaic expressions are also less then watertight: they are more frequent in some situations than in others. For example, **Scheibman** (volume 2) shows how the use of the expression *for me* can have different pragmatic functions within discourse. Aspects of language that resist absolute, non-statistical characterizations have not been in the forefront of typical linguistic research. **Pawley** (volume 1) discusses the place different models of language assign to speech formulae, which he suggests are, along with phrasal lexical units, the main building blocks of connected speech and play a key role in linguistic competence.

In contrast to generative approaches, usage-based approaches have attributed a much more prominent role to formulae. Bybee (2006: 711) states: "A usage-based view takes grammar to be the cognitive organization of one's experience with language." In volume 1, **Bybee and Cacoullos** suggest that frequency of use is a major determinant of the rate at which a multi-word unit or construction grammaticizes over time. **Bannard and Lieven** (volume 2) note that in usage-based theories, novel utterances are produced and understood by analogy with previously experienced language, while in generative theories, productivity comes about because of "some language specific, pre-experiential mechanism such as innate linking rules." They claim that language is learned both by observing and by interacting with others and that reuse of language is the basis for communication. **Peters** (volume 2) also emphasizes the role of experience as the basis for children's eventual construction of internal representations of the language they hear. **Erman** (volume 2) claims that particular types of collocations "reflect language users' experience as social beings." Other usage-based explanations include how people learn the semantic prosody of verbs (**Ellis and Frey,** volume 2), how L2 learners acquire Japanese tense-aspect markers (**Sugaya and Shirai,** volume 2), and the content of historical metaphors about the spleen (**Mischler,** volume 1)**.**

Given that it is important to study formulae, what is it that needs to be learnt about them? Here are some research questions.

(1) Structure and distribution:
- What **structures** are used in formulae in a given language and across languages?
- What **meanings** are expressed formulaically in a given language and across languages?
- What is the **distribution** of common forms and meanings across dialects, speech styles, and languages?

(2) Historical change:
- How do formulae arise?
- How do formulae change in the course of history?

(3) Acquisition and loss:
- How are formulae acquired and used by children learning their first language?
- How are formulae acquired by second-language learners?
- How are formulae retained, altered, or abandoned in geriatric and pathological cases?

(4) Psychological reality:
- How are formulae stored and processed by the mind?
- What is the relationship between formulaic patterns and thought patterns?

(5) Explanations:
- Why are the facts about the structure, distribution, individual and historical change and psychological reality of formulae the way they are?
- Why are there formulae in human languages at all?

In what follows, we will survey the papers of this collection from the point of view of how they address the five main headings given above. Several of the papers address more than one issue and thus this survey may refer to them more than once. However, in the book itself, we classified the papers according to their strongest focus.

3. Synopsis of both volumes

3.1 Structure and distribution

One question surrounding formulae concerns the types of structures that are used in formulae and the meanings that they express. Authors in the current book examine many different types of formulaic structures including grammatical constructions, idioms, collocations, and compounds. **Calude** (volume 1) discusses a particular subtype of English cleft constructions dubbed demonstrative clefts. Examples are *that's what I said*, *that's why I object*. She demonstrates four formulaic characteristics of this construction: structural fixedness, fluent (cohesive) phonological shape, the non-salient (vague) reference of the demonstrative involved, and prominent frequency in informal, conversational English.

Szerszunowicz (volume 1) analyses Polish and Italian idioms that include place names that have evaluative connotations, such as English "The Boondocks". These toponyms stand as symbols of a given culture and are by and large untranslatable from one language to another.

Two papers focus on collocations. **Erman** (volume 2) examines collocations that have fused meanings in the written essays of learners of English. **Ellis and Frey** (volume 2) use an affective priming task to examine the semantic prosody of a set of English collocations.

Haiman and Ourn (volume 2) describe formulae in Khmer of a special structural type: symmetrical compounds, similar to English *last and final*, or *pel-mell*. In Khmer, they occur both in ritual language and in everyday conversation.

A second question concerns how formulae are distributed. A number of papers in this book focus on the use of formulae in particular genres. Two of the papers study formulaic expressions in scientific discourse across academic disciplines. **Dorgeloh and Wanner** (volume 2) survey the use of expressions like *This paper argues …* or *This article analyzes …* , which constitutes one of four different reporting styles they have identified in scientific papers and abstracts in particular. They find that the "*paper* construction" is most prevalent in the humanities literature. The paper by **Kerz and Haas** (volume 1) is related in topic but broader in scope: the authors study the function of prefabricated chunks of various sorts in academic discourse, such as *The aim is to analyze …* or *The survey shows …* These expressions are shown to marks specific stages of the research process reported on.

Sams (volume 1) looks at varying degrees of formulaicity and argues that genre dictates the degree of quotative formulaicity, both in specific lexical choices and constructional patterns. She argues that fiction writing is more likely to depend on the use of null quotatives, adverbs or adverbial phrases or clauses, and pronominal speakers, whereas newspapers are more likely to depend on quoting verbs in the communication/statement frame, initial quotatives, inverted quotatives, and adjectival phrases or clauses. The dependence on these features closely relates to the function of each of the genres.

Gruber (volume 2) also describes a specialized genre, criminal defendants' use of a particular type of formulaic language (acceptance of responsibility) during sentencing hearings.

Thompson and Ono (volume 1) argue for a usage-based approach to reveal that interactional and cognitive practices are deeply intertwined in the lexical category of adjectives for Japanese speakers. They show that adjective usage in conversation is intricately bound up with fixedness and frequency and argue that "learnt as a chunk" plays a much larger role in the use of adjectives in Japanese than has been assumed in the literature.

3.2 Historical change

Wray (volume 1) suggests that formulaic status may protect a word string from language changes. A formula may retain its meaning over time even as the grammatical rules of the language change and, as a result, a string that was originally analyzable can become opaque.

Peters (volume 2) suggests that the same elements that create change in child language also operate to produce historical changes. Specifically, those elements

in adult language that have looser or minimal connections with other elements in the system are the ones that grow and change. Some of them eventually are grammaticalized. **Bybee and Cacoullos** (volume 1) examine the role of formulae in the diachronic development of *can* in English and "*estar* + gerund" in Spanish, arguing that formulae contribute to the process of grammaticization by demoting the independent lexical status of the parts and promoting the productivity of the construction.

Mischler (volume 1) explores historical metaphors in English centering on the human spleen. He suggests that a particular cultural model (the Four Humors model of medicine) accounts for specific characteristics of spleen metaphors. He notes that particular historical cultural models can account for certain conceptual metaphors and how they change over time.

Lindquist (volume 1) uses data from the British National Corpus to examine how formulae involving prepositions and body parts become lexicalized and acquire more abstract, metaphorical meanings.

Lancioni (volume 1) analyzes certain grammatical features in Arabic, which he argues have a formulaic origin. His analysis focuses on the formulaic features in Classical Arabic and Modern Standard Arabic which are missing from spoken Arabic variants; these features range from text chunks to morphological and syntactic patterns (including redundant case affixes, and syntactically determined partial agreement). The general consequence of his hypothesis is that formulaicity in written languages can be strongly reinforced by the model of literary varieties, even long after the original textual constraints disappear. He argues that the influence of Modern Standard Arabic on modern spoken varieties shows the possibility that such formulaic features find their path through spoken languages.

Wilson (volume 1) examines the diachronic development of exemplar clusters, showing how certain formulae that use a verb of becoming + adjective serve as central members of exemplar categories and how the members of these categories mutate over time.

3.3 Acquisition and loss

A number of authors claim that formulaic language is the starting point for **first-language acquisition. Bannard and Lieven, Peters** (both in volume 2), and **Wray** (volume 1) all agree that development proceeds from formulaic language to analyzed forms rather than vice versa. **Wray** (page 32) suggests that the learner "attempts to map the largest possible form onto a reliable meaning." If there is no need for further analysis, the chunk will remain unanalyzed. When the learner encounters variation within a recurrent pattern, s/he will figure out where the variation is and keep the remainder fixed. That is, the child begins with multi-word

strings and over time analyzes them into smaller components on a "needs only" basis. Lexicons reflect patterns of variation in the input.

Bannard and Lieven and **Peters** trace the analysis of recurrent patterns during language acquisition. According to **Bannard and Lieven** (volume 2), the basic sequence is that adults produce many item-based phrases such as *Where's x?* when they speak to young children. Children analyze these chunks and eventually develop more general categories or schemas such as a transitive construction. They then connect their constructions into complex networks. **Peters** (volume 2) describes how children begin with unanalyzed chunks and discover how they relate to one another, resulting in a gradual shift from unrelated items to a system of related items. The process can be traced by examining occasions where the child's use deviates from adult analyses.

Kurumada (volume 2) shows how the Japanese *wa* + NP construction is very frequent in mother-child interaction. It is acquired early by children and is an important tool in learning new vocabulary.

The acquisition of formulae presents a special problem for **second-language learners**: they have to get them "just right" both in form and in use. An example is the English formula *Have a nice day!* It admits some lexical variation, such as *Have a good day!* or *Have a great day!* but the form *Have great days!* used as the parting phrase in an e-mail message by a Korean student is off the mark. **Erman** (volume 2) suggests that learning formulae is problematic for second language learners because, compared to first language learners who usually hear formulae repeatedly, second language learners have less extensive language exposure. She examines different types of formulae used in the written compositions of university students who are native English speakers compared to those who are learning English. She finds that the learners underuse collocations, which makes their compositions appear less native-like.

In his paper on the acquisition and use of formulae by learners of English as a Second Language, **Ohlrogge** (volume 2) addresses two questions, one about the kinds of formulae used by intermediate-level learners in high-stakes written exam papers, the other about formulaic expressions used by high-scoring and low-scoring learners. He finds eight subtypes of formulae in the exams of the intermediate-level learners and finds some differences depending on the scores of the students.

Sugaya and Shirai (volume 2) suggest that the early acquisition of Japanese tense-aspect morphology by L2 learners shows verb-specific patterns and that the learners gradually attain productive control of tense-aspect forms, which is consistent with the proposed developmental sequence: formula > low-scope pattern > construction (Tomasello 2003; N. Ellis 2002). These findings are similar to those of **Bannard and Lieven** (volume 2) in first language acquisition.

Rott (volume 2) examines how awareness-raising tasks can be used to facilitate the acquisition of formulae in L2, finding different degrees of effectiveness based upon the genre.

Finally, in the area of language loss, **Van Lancker Sidtis** (volume 2) examines evidence that the comprehension and production of formulae is preserved in patients with left hemisphere damage but lost or impaired in those with right hemisphere or subcortical damage.

3.4 Psychological reality

Wray's working definition of formulaic sequences (2002: 9) includes the notion that these structures are stored and retrieved from memory as wholes. In volume 1 of this book, **Wray** (endnote 1) argues that, while this may be true, there is no independent way to determine whether something is or is not stored or retrieved as a whole. She suggests that experimental methods cannot establish whether an individual is actually exhibiting "holistic access or fast-route componential decoding."

Nevertheless, a number of authors in the book argue for the psychological reality of formulae as wholes. **Bannard and Lieven** (volume 2) review experimental work that they believe provides evidence that multi-word utterances can be stored as a whole. They cite research into the statistics of natural languages that has shown mathematically that the most efficient way (i.e., requiring the fewest processing steps) to understand or produce language is to have information stored in memory in a redundant manner. For example, an adult might store *what's that* as a unit even though s/he knows that it is related to *what is that*. **Kapatsinski and Radicke** (volume 2) examine the effect of word frequency and phrase frequency on the speed of detection of word parts, and their results support the hypothesis that high-frequency formulae are stored in the lexicon in the same way as words are.

Ellis and Frey (volume 2) are interested in the psychological reality of semantic prosody and collocation. They show that verbs that are strongly positive or negative in semantic prosody show affective priming. That is, participants in their experiments were quicker and more accurate in deciding that a target word was generally positive (pleasant) or negative (unpleasant) if it was preceded by a prime that matched in semantic prosody. Their results support the psychological reality of semantic prosody at the semantic access stage of lexical processing.

Van Lancker Sidtis (volume 2) argues for the use of a dual process model of language, in which the holistic mode is used to process formulae while the analytic mode is used to generate new and creative utterances. These two modes also interact with one another when processing schemata, or fixed forms with one or more open slots.

3.5 Explanations

Why are there formulae in human languages?

As noted in several of the papers mentioned in section 3.2 above, the engine that drives the genesis of formulae is grammaticalization: the process of phonetic simplification and semantic bleaching that also underlies the origin of grammatical markers.

But what drives the grammaticalization of ordinary phrases into formulae? **Bannard and Lieven** (volume 2) argue that formulaic language occurs because of a basic law of psychology: humans show preferences for things they have experienced previously. Examples they cite include the fact that humans link to web sites they have used before, they cite papers they have cited before, and they use words and constructions that have been used previously. They point out that the likelihood of a word being repeated depends on how often it has been encountered before.

Several authors make the point that there is a trade-off between the ease of processing of formulaic utterances and the flexibility provided by novel utterances. One example is described in **Wray's** paper (volume 1). When people use augmentative communication (devices designed to support the communication of individuals who are unable to use oral speech) to type in anticipated language structures in advance, their savings in processing speed are offset by their inability to tailor their messages to individual circumstances during an actual conversational interchange. Another study which highlights the processing advantage of formulaic utterances is **Iwasaki**'s paper on "time management expressions" in English and Thai (volume 2), such as English *you know* and *I mean*. He suggests that these expressions serve as aids to the speaker in the difficult task of having to transfer ideas and images into linguistic form. Since the speaker must both think and speak concurrently, such formulae gain time for him. Yet another example is described by **Gruber** (volume 2), where criminal defendants' use of formulaic language such as "I accept responsibility for what I have done" can be interpreted as acceptance of criminal status and remorse, but can also make the criminal appear insincere. Use of novel language in accepting responsibility, such as "I know I did this to myself" can make the defendant appear more sincere, but may signal that s/he is less willing to accept the social role identity of criminal.

Formulae can serve many functions including the identification of different types of genre, the introduction of new vocabulary, and various pragmatic and aesthetic functions. In their survey of the use of expressions like *This paper argues …* or *This article analyzes …* , **Dorgeloh and Wanner** (volume 2) find that the function of the "*paper* construction" is to emphasize the argument-constructing nature of a paper as opposed to fact-reporting articles.

A different kind of function is evident in the case of the Japanese *wa*-construction as it occurs in mother-child interaction. **Kurumada** (volume 2) suggests that *wa*-plus-noun sequences provide an ideal context for the mother to introduce new vocabulary to the child and for the child to ask questions about the names of unfamiliar objects.

Scheibman (volume 2) focuses on the pragmatic functions of formulae within discourse, such as marking an evaluative speaker or making polite requests.

In their paper on Khmer symmetrical compounds, **Haiman and Ourn** (volume 2) argue that formulae may have purely decorative functions satisfying æsthetic desiderata of the interlocutors and may have been created not for their meaning but for their phonetic characteristics. The æsthetic virtue of these expressions is parallelism of structure, which, as they point out, is also evidenced in some instances of grammatical agreement, reduplication, structural priming and even baby talk. They cite analogous, æsthetic formulae from several other languages as well.

4. Conclusions

As in other aspects of the study of human cognition and social behavior, a central question is the balance of freedom and constraint: given that there is a system consisting of rules, how much freedom are we nonetheless allowed? Formulae are distinguished from ordinary sentences exactly by the limitedness of structural and lexical choices.

For this reason, the existence of formulae in language bears on a central question of linguistic description. Similar to the description of any complex object outside of language, a basic issue in linguistics is one of segmentation: what units should be posited to facilitate the formulation of maximally fruitful generalizations (cf. Aronoff 2007)? Some of the units that have multiply proven their significance in linguistic analysis are sentences, clauses, phrases, words, morphemes, syllables and sounds. That entire constructions must also serve as basic units of linguistic description has been highlighted by work on construction grammar (Goldberg 1995; Croft 2001). Formulae are a special type of entity: they are rule-governed in form and may even be compositional; but they manifest only one – or only a few – of the various formal structures that the language allows for the expression of their meaning. Thus, despite their being phrase-size or sentence-size, and even though they may be subjected to further partonomic analysis, formulae must be assumed to be one of the basic units of linguistic description.

Linguistic formulae are not unparalleled outside language. Frequently performed routines such as playing a favorite piano piece, starting a car, brushing one's teeth, or even walking are akin to linguistic formulae in that they, too, form unified

chunks of behavior. A seemingly paradoxical feature of such behavioral chunks is that while they may be conceptualized as single wholes, under certain conditions, users can also readily analyze them into components. People may alternate between the two viewpoints or even keep both in mind at the same time.

The paradox of something being both one and many, however, is apparent only: a conceptual tool fundamental to human cognition – whole-part relations – resolves it. Given that we conceive of wholes consisting of parts, we can view "one" as being "many" and "many" as being "one" without inconsistency. Formulae and other chunks of routinized behavior are distinguished by the tenuous balance between the holistic and analytic view being shifted in favor of the holistic viewpoint.

References

Aronoff, Mark. 2007. In the beginning was the word. *Language* 83(4): 803–630.

Bybee, Joan. 2006. From usage to grammar. The mind's response to repetition. *Language* 82(4): 711–733.

Cameron-Faulkner, Thea, Elena Lieven & Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science* 27(6): 843–873.

Corrigan, Roberta. 2004. The acquisition of word connotations: Asking 'What happened?' *Journal of Child Language* 31: 381–398.

Croft, William. 2001. *Radical construction grammar. Syntactic theory in typological perspective.* Oxford: OUP.

Ellis, Nick. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition* 24(2): 143–188.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text,* 20(1): 29–62.

Goldberg, Adele E. 1995. *Constructions: a construction grammar approach to argument structure.* Chicago IL: University of Chicago Press.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3): 491–538.

Sinclair, John.1991. *Corpus, concordance, collocation.* Oxford: OUP.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge MA: Harvard University Press.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.

PART I

What is Formulaic Language

Grammarians' languages versus humanists' languages and the place of speech act formulas in models of linguistic competence

Andrew Pawley

> [For the interpreter of texts] language appears to begin where analytical grammar leaves off. (Gregory Schrempp 1992: xvii)
>
> There continues to be a need for a model of natural discourse that pulls together the diverse cognitive and social factors responsible for the shape of language. (Wallace Chafe 1996: 49)

Abstract

The paper begins by observing that the notion of what a language consists of is problematical, reflected in one scholar's comment that, for the interpreter of texts, "language appears to begin where analytical grammar leaves off". Section 2 describes 'speech act formulas' as conventional bundles of eight or nine different features, including several that are not normally considered part of grammar or lexical items, such as discourse function, idiomaticity constraints, special 'musical' features such as voice quality and volume, and body language. Section 3, viewing the period before 1970, asks what place was given to speech formulas in analytic grammars, on the one hand, and in treatments of language by conventional lexicographers and other humanists (scholars chiefly interested in language as an expression of human affairs), on the other. Section 4 comments on the circumstances

that led some linguists and lexicographers in the 1970s to treat speech formulas as playing a central rather than a peripheral role in linguistic competence. The final section asks whether in recent decades 'usage-based' or 'discourse-based' approaches, often drawing on large electronic corpora, have led linguists to modify their views of what languages are, or whether most are still working with a grammar-and-lexicon centric model. It seems that certain methodological and theoretical biases act as conservative forces, keeping linguists focused on grammatical and lexical form, while paying relatively little attention to the full complexity of speech formulas and their role in fluency, idiomaticity, coherence, appropriateness, wit and other highly-valued facets of ordinary language use.

1. Languages as problematical constructs

This paper will reflect on the place different models of language give to speech formulas, which I suggest are, along with phrasal lexical units, the main building blocks of connected speech and play a key role in linguistic competence.[1]

It is perhaps not sufficiently appreciated in linguistics that 'languages' are problematical constructs. I refer here in particular to the kinds of knowledge to be considered part of a language, to the scope and content of linguistic competence.[2] It is generally agreed that languages are codes for linking signs and meanings but this broad definition is problematical because one can consider more or fewer such linkages to be properly part of a language. Different groups of scholars adopt different definitions according to their particular theoretical perspective or customary concerns. Surprisingly, textbooks seldom acknowledge or reflect on these differences.

During the 19th and 20th century, two major scholarly traditions, those of grammarians and lexicographers, dominated the descriptive analysis of language and their ideas about what a language is differed in certain important respects.[3]

---

1. I have profited from helpful discussions of some of the topics treated here with Wally Chafe, Tony Cowie, Frank Lichtenberk, Jim Miller, Mickey Noonan, Malcolm Ross and Alison Wray. I have been much influenced by the ideas of George Grace, beginning with his course in ethnolinguistics at the University of Hawaii in 1975; the basis of Grace (1981), where he says "The key problem in understanding how language works is that of understanding how it is used to say things" (1981: 35).

2. I do not refer here to the question of whether we should regard languages as a social contract deposited in the brain of each individual (as Saussure put it) or something belonging to the collective minds of a community, or, as externally observable patterns of speech or text (or all of the above), or to when a set of speech traditions are to be considered 'dialects' or 'varieties' rather than separate languages.

3. Newmeyer (1986) makes a somewhat similar distinction between structuralist and humanist treatments of language.

The grammarians' tradition has its origins in pedagogical and philosophical grammar, going back to Classical Greece and Rome, but it gained a much sharper edge with the development of synchronic linguistic theory in the 20th century.[4] The grammar-based tradition underlies much of modern linguistic theory as well as descriptive practice. The lexicographers' approach owes more to what I will call the humanist tradition to the concerns of people chiefly interested in language as an expression of human affairs, such as lexicographers, translators, lawyers, dramatists, novelists, philologists, cultural anthropologists, clinical psychologists and aphasiologists, to name just a few.

Models of language used by grammarians tend to be rigorously defined but highly reductionist. The task of describing a language is equated, essentially, with writing a grammar and a lexicon (or a grammar operating on a lexicon). Central to grammarians' work is the distinction between 'well-formed' and 'ill-formed' form-meaning pairings, which in turn are equated narrowly with 'grammatical' versus 'ungrammatical' pairings. A grammar is an algorithm for specifying all and only the well-formed pairings of form and meaning together with their structure. It is not the business of a grammarian to consider what such pairings are used for. Grammarians stress the autonomy of the core components of language, as self-contained systems standing apart from social context and use.

Humanists seldom provide explicit models of language. However, from their practices, well exemplified in lexicography, for instance, it is clear that their perspective differs from that of grammarians in at least two important respects. First, they are much concerned with what languages *do*. Humanists regard languages as instruments for saying particular things for particular communicative and social purposes. Second, this perspective leads them to take a much broader view of what languages *consist of*. While grammaticality has a place, there are other concepts that are central to understanding how languages function in human affairs and to what it takes to know a language. In this respect, languages are not autonomous systems. One could say that 'well-formedness' has a much broader significance for humanists. No doubt the anthropologist and folklorist Gregory Schrempp had these things in mind when he wrote that, for the interpreter of texts, "language appears to begin where analytical grammar leaves off" (Schrempp 1992: vii).

It might be thought that the grammarians' and humanists' views of language complement one another, with the humanists' view taking for granted the core components that engage grammarians but also placing great importance on cer-

4. One could call this the 'structuralist' tradition but the latter term might be misread as excluding generative grammar.

tain kinds of knowledge or behavior that grammarians regard as peripheral. That is, the grammarians' account of language can simply be extended to accommodate the peripheral bits, without affecting our ideas of how the core works.

There is an element of truth in this idea. However, I will argue that the situation is more complicated. The 'peripheral' bits have turned out to be much more pervasive than was once generally thought and paying close attention to the peripheral bits may force changes to our understanding of how the whole system works.

Section 2 of this paper will briefly review the characteristics of 'speech act formulas'. This large class of expressions presents a number of challenges to anyone seeking to describe a language or to define linguistic competence, challenges broadly representative of the class of speech formulas as a whole. Section 3 compares grammarians' and lexicographers' views of what a language is and asks what place speech formulas were given in analytic grammars, in lexicography, and in other treatments of language in the humanist tradition before about 1970. Section 4 comments on the circumstances that led various linguists and lexicographers in the 1970s to pay more attention to formulaic language and it reviews some research projects of the 1970s and 80s that treated speech formulas as playing a central rather than a peripheral role in linguistic competence. Finally, in section 5, I will ask whether extensive research using 'usage-based' or 'discourse-based' approaches in recent decades has led linguists to modify their views of what languages are, in the direction of the humanist view, or whether most linguists are still working with a grammar-based model.

2. Speech act formulas

Following Austin (1962) and Searle (1969), I use 'speech act' broadly to refer to any utterance that performs a discourse function over and above those of referring and predicating. While this definition is perhaps dangerously general, it is intended to refer to prototypical discourse functions such as greeting, welcoming, farewelling, introducing, complimenting, insulting, apologising, and so on. Speech acts are often performed using a conventional form of words, a formulaic expression, that is bound to a particular discourse context and a particular discourse function. Some formulaic expressions are single word interjections or phrases, e.g., *Hello, Sorry, Congratulations, Giddyup, Excuse me, Yours sincerely, on that note, on the contrary*. But a great many are clause-sized, e.g., *You can say that again*, *Long live the Queen*, *I'd like to propose a toast to our host, I'm sorry to keep you waiting*, *I wouldn't dream of it, Go to hell,* and *That's another story.*

In grammatical terms, a *productive formula* is a construction type that is partly lexically specified and so can generate a number of formulaic expressions that belong to the same family. Productive speech act formulas differ from both

typical ('word-like') lexical units and abstract grammatical constructions. A typical lexical unit is, minimally, a bundle of perhaps four features: form, meaning, grammatical category and marking for register. An abstract (or purely syntactic) grammatical construction is a formula in the notational sense but it contains no specific lexical forms or meanings and is not tied to specific discourse functions or contexts. A typical speech act formula, by contrast, is a bundle of some eight or nine different features.

a. discourse function
b. discourse context
c. literal meaning
d. pragmatic (meaning more specific than discourse function)
e. regular grammatical structure and constraints inherited from a major construction
f. idiomaticity constraints
g. basic phonology inherited from general phonology
h. music – speech act formulas require a certain intonation and prosodic pattern, and sometimes a specific volume level, voice quality, and the like.
i. body language – gesture, posture, eye gaze, etc.

Here is a very brief sketch of a particular speech act formula in these terms (omitting (e) and (g)).

(4) **(I'm) (INTENSIFIER) PLEASED to meet you!**

Discourse context. A response move in a first meeting with the addressee, after the addressee has introduced himself or been introduced. Near functional equivalent in this context is *How do you do?*, but *(I'm) pleased to meet you* is warmer.
Discourse function. To warmly acknowledge the addressee's status as a new acquaintance.
Linguistic meaning. Literal (i.e., not an idiom).
Music. Should be spoken in a friendly manner, with a bright tone. There should be a main stress on *pleased* and *meet* (not on *I'm* or *you)*.

Idiomaticity constraints:

a. **Grammatical structure.** Must be an assertion in the present tense, as above, e.g., cannot be questioned or negated or put in another tense without destroying the formula. (The formula *be-TENSE* NP_i *pleased to see* NP_j *!* as in *Are we pleased to SEE YOU?!,* has a distinct discourse context and function.) In its syntactic form this formula belongs to a class of constructions that consists of subject + copula/quasi-copula + adjective of emotion + infinitival complement (to V + O), e.g., *She was relieved to find us, They seemed pleased to meet us,* or *I'd be delighted to go*, but it does not inherit the general characteristics of this class.

b. **Lexical variability.** The unmarked adjective is *pleased* but any of a few others, e.g., *delighted*, *honoured, glad, thrilled*, can be substituted though with certain contextual conditions. The adjective can be modified by certain intensifying adverbs, e.g., *very*, *really* or *so*. However, there are subtle constraints and nuances associated with the use of these intensifiers, as there are with the marked adjectives.

c. Body language. Ideally the speaker and addressee should be facing each other, should make eye contact as the greeting is spoken and should be more or less stationary (not walking away from each other). Unless physical circumstances make it awkward it is customary to offer a handshake either during, or in the seconds before or after uttering the formula.

A speech act formula is a social institution: it specifies, in more or less detail, *what* may be said (meaning), *how* it may be said idiomatically (form), *why* it is said (function) and *when* it may be said (context). To deal with speech act formulas satisfactorily we must invest heavily in the analysis of the social conventions that govern discourse. That is, this work is best done as part of a broader study of discourse structure and this in turn is best done as part of a study of social actions and norms.

The number of productive speech act formulas known to the ordinary mature native speaker of English perhaps runs into the thousands. The number of lexically specific formulaic expressions realising these is indefinitely large.[5]

3. Grammarians' languages, humanists' languages and the treatment of speech formulas

Let me now say more about the prevalent conceptions of what languages are, among grammarians and among lexicographers and other humanists, as these developed up to about 1970. I do not suggest that every grammarian or humanist held these views. To make generalisations one must use a broad brush.

3.1 Grammarians' languages

It is no surprise that clear statements of the grammarians' view are plentiful in the early literature on generative grammar. Thus, Seuren (1969: 3) writes that "a

5. Not all speech formulas are speech act formulas. And many speech acts can be done using a non-formulaic form of words, deriving speech act status from the conventions of conversational implicature. For example, the highly productive time-telling formula *It/The time be-TENSE M to/past H*, is not bound to a speech act function but it can be, and often is pressed into service as a speech act, e.g., as a signal to start, to hurry, as a complaint, and so on.

grammar, or grammatical description, is essentially a device for defining a language". This echoes Chomsky (1965: 4) who writes "A grammar of a language purports to be a description of the ideal speaker-hearer's intrinsic competence". Lexicon is seen as part of grammar, though its precise manner of interaction with grammatical rules has been a matter of some debate. For present purposes the key point is that the grammarians' notion of the lexicon is a minimalist one, in that it excludes all well-formed pairings of form and meaning, i.e., those that can be generated by rules of grammar.[6]

Grammarians tend to be admirably egalitarian in two respects. Firstly, all sentences, indeed all well-formed strings, are equal in the sight of grammar. A much-cited proverb, a standard form of words for performing an apology, a compliment or a marriage ceremony have no more status than any nonce sentence. It does not matter if a particular sentence (or phrase) has never been uttered, or if it is semantically anomalous (the famous 'colourless green ideas sleep furiously'); what matters is that it is grammatical.

Secondly, grammarians do not take kindly to arguments that some languages are inferior to others in their expressive power. The flexible, expressive potential of grammar makes it possible for individual language users – in principle – to talk about any conceivable idea or subject matter, however clumsily. In the Chomskyan paradigm of the 1960s the emphasis was strongly on the power of syntax to create novel sentences. This emphasis on novelty was extended from competence to performance: "It is evident that rote recall is a factor of minute importance in ordinary use of language" (Chomsky 1964: 914).

It has been suggested to me that the equation of languages with grammars (more precisely, with phonology and grammar-including-lexicon) belongs to a rather short phase in the history of linguistics, associated with the dominance of generative grammar in the 30 years or so after the publication of Chomsky's *Syntactic Structures* in 1957. This is not the case. While the statements quoted above reflect the conceptual framework and metalanguage that Chomsky brought to linguistics in the 1950s and 60s, they have clear antecedents in the writings of

---

6. Among generative grammarians conceptions of the lexicon have changed a good deal since the early days, when scholars were extremely optimistic about the power of syntactic rules to generate complex words. At least since Chomsky's "Remarks on nominalization" (1970) this optimism has been tempered by the realisation of certain difficulties with minimalist lexicons and the scope of the lexicon has been gradually extended. Even so, I believe that in many quarters the extensions have been made grudgingly, without giving up the basic principles underlying the grammar-lexicon model, namely well-formedness and economy of description.

earlier generations.[7] The successful development of the comparative method of historical linguistics in the 19th century was based on the discovery that sound change is systematic and independent of the speech community's physical environment, social habits and moral values. Arguably the most influential book in linguistic theory in the first half of the 20th century was Saussure's *Course in General Linguistics* (1916), which defined the object of enquiry of structural linguistics. As Robbins points out in his *History of Linguistics*, Saussure was at pains to emphasize, first, that there needs to be a synchronic linguistics whose object of study is languages as self-contained systems of communication existing at a particular point in time; second, that while *parole*, or speech, provides the raw data, "the linguist's proper object is the *langue* (linguistic competence) of each community, the lexicon, grammar and phonology, implanted in each individual by his upbringing in society" (Robbins 1967: 200); and third, that lexical, grammatical and phonological elements are to be defined by their place in the system.

A bit later came Bloomfield's manifesto "A set of postulates for the science of language" (1926) and the 2nd, much revised edition of his influential book *Language* (1933). Bloomfield defined a language as "the totality of utterances that can be made in a speech community" (1926: 154), where an utterance consists of one or more sentences. Grammatical regularities are central and "the lexicon is really an appendix of the grammar, a list of basic irregularities …." (Bloomfield 1933: 274).

Among the clearest presentations of the 'Neo-Bloomfieldian' tradition that dominated American structural linguistics from the 1930s to the 1950s is that given in Charles Hockett's *A Course in Modern Linguistics* (1958). Hockett says that a language consists of five main subsystems, three central, two peripheral (1958: 137–8). The central subsystems are the grammatical, phonological and morphophonemic systems, central because they have nothing to do with the nonspeech world. The peripheral systems are those of semantics and phonetics. Semantics is peripheral because it impinges on the physical and social world, as well as on grammar. Phonetics is peripheral for other reasons.

The rationale given for not dealing with facts such as the pragmatic functions or social status of particular forms was sometimes the assertion that language use is not part of linguistic competence, at other times the belief that the time is not yet

7. Statements of belief are not the only measure of one's world view. Scholars must be judged not just on what they profess to believe but on what they do and get rewarded for doing. For centuries descriptions of languages have consisted of grammars and dictionaries. In the modern era the overwhelming emphasis in descriptive and theoretical linguistics has been on producing grammars and on the theory of phonology, morphology, syntax and structural semantics. Pragmatics, for example, was a latecomer.

ripe. Thus, Lyons (1968) followed Chomsky in delimiting the scope of linguistic theory thus:

> … linguistic theory, at the present time at least, is not, and cannot, be concerned with the production and understanding of utterances in their actual situations of use … but with the structure of sentences considered in abstraction from the situations in which actual utterances occur. (Lyons 1968: 98)

Of course, to do science it is necessary to be a reductionist, for practical reasons: one must define an object of study that is manageable. In seeking to establish a scientific linguistics it was legitimate and sensible for structural linguists to push to one side those conventions of communicative behaviour that seemed less central, or less amenable to systematic analysis, than those of phonology and grammar. What is problematic about the statements quoted above, obviously, is the idea that *languages* are to be equated with *grammars* and that *linguistic theory* should be chiefly about *grammar*.

On that note, let us turn now to the treatment of formulaic expressions by grammarians before 1970.

Grammarians acknowledged their existence but gave them short shrift. Speech act formulas are generally classed under, or treated together with 'minor constructions', and given a page or two with some notes on their discourse functions. In *Language*, Bloomfield gives a page (1933: 176–7) to what he calls "minor sentences". He distinguishes three types:

1. The completive type, which "supplements a situation". This type mainly consists of truncated answers to questions, e.g., *Yes, No, With whom?, When? Tomorrow morning.*
2. The exclamatory type, which he says occurs "under a violent stimulus" – meaning I think either physiological or social. Examples are *Ouch, Damn it, This way please, Hello, John*.
3. The aphoristic type, e.g., *The more the merrier*, *First come, first served*.

Forty years on, in their *A Grammar of Contemporary English*, the most comprehensive English grammar to date, Quirk et al. (1972) give just three pages out of 1100 to the discussion of what they call "formulaic utterances, greetings, etc." They place these formulaic utterances among the "residue of minor classes" of utterance which are "something of a museum of oddments" (p. 411). Quirk et al distinguish four "minor utterance classes".

1. Minor constructions of a sort that "enter few of the relations of substitutability that are common to one of the major classes", e.g., *How do you do?, Why get upset?*, *How about joining us?*, *To think I was once a millionaire.* They call these

'formulae' and distinguish them from syntactically more standard minor constructions, though they allow that there is a gradient.
2. Aphoristic sayings: *Least said, soonest mended.*
3. Interjections: emotive words with no referential content.
4. Greetings and other formulas used for stereotyped communicative situations. They distinguish 13 main types: greetings, farewells, introductions, reactions signals, thanks, toasts, seasonal greetings, slogans, alarm calls, warnings, apologies, imprecations, expletives, and miscellaneous exclamations. Most such formulas are grammatically irregular or defective.

From a grammarian's standpoint the description of these types as "a museum of oddments" is accurate. But there are indications that Quirk et al. underrate the importance of formulas in ordinary language. Three pages out of 1,110 is one indication. Then, too, their list of 13 major types barely scratches the surface of speech act types. And to describe formulas as "used for stereotyped communicative situations" (412) is true but also serves as something of a put-down. Instead of 'stereotyped' one might use 'structured' and point out that most of language use consists of structured communicative situations and it is such structures that are the basis for word play and much other creative use of language. Lyons (1968: 98, 177–8) makes some insightful general remarks about what he calls "situation-bound", and "ready-made expressions" but he slips when he says these locutions make up a relatively small class.

3.2 Humanists' languages

Humanists are not linguistic egalitarians. Any language or linguistic genre evolves as a particular community's means for talking about particular subject matters and as a component of other culturally-authorised activities. For these purposes its speakers need a large repertoire of conventional expressions, including word-level expressions for concepts that are significant in the culture and sentence-level expressions that do particular jobs in discourse and social life. Humanists tend to focus on differences in the expressive resources of languages, on differences in what can be said, or what it is appropriate to say, and on ways of saying it, and find these differences endlessly fascinating.

A convenient place to find clues to the humanist view of language is in conventional dictionaries, both general and specialised. Lexicographers practice a craft that has evolved by doing rather than theory-building. Although their practices are not always completely systematic or consistent they are consistent enough to reveal a fairly coherent view of what counts as a lexicalised expression. There is a common membership in the ideal lexicons of lexicographer and grammarian, namely, the

form-meaning units that are either unanalysable or irregularly formed. Where they diverge markedly is in the treatment of complex expressions that are well-formed. Lexicography has always been usage-based, not grammaticality-based.[7]

A look through the pages of any large general dictionary of a European language will show that many literal expressions, i.e., well-formed form-meaning pairings, are included. For example, among the compounds listed under *door* in the *Shorter Oxford* are *door alarm*, *door frame*, *door mat*, *door post*, *door step* and *door stop*, in what seem to be regular senses: 'alarm for a door', 'frame for a door', etc. Under *blood*, the 2nd edition of *Webster's New World English Dictionary* lists such well-formed compounds as *blood-colored*, *bloodstained*, *blood test*, *blood type* and *bloody-faced*, analogous with thousands of other possible compounds of the form *X-colored*, *X-stained*, *X test, X type*, *Xy-faced.* Beside *forget*, Webster's gives *forgetter*, *forgettable, forgettableness,* among other derivatives, defined in their literal senses. The suffixes *–er*, *-able* and *-ness* are extremely productive.

What is going on here? It is noteworthy that none of the dictionaries list compounds that are merely possible expressions. For instance, we don't find entries for *table alarm*, *table step*, *grass-colored, grass-stained*, or *grass test*. It seems that the lexical status of a composite expression is determined with the following questions in mind: (1) Is the meaning a conventional concept, one familiar to members of the speech community? (2) If so, is the form in question the standard way (or a standard way) of expressing that concept? We might say that (1) and (2), taken together, constitute the *standard usage* or *conventional usage principle*: any highly conventional form-meaning pairing is a lexicographer's lexeme.

What exactly makes an expression 'standard' or 'conventional'? Frequency of use is certainly one ingredient. But frequency is not the whole story. To say that a word or phrase has conventional or standard status is to say more than that it recurs in speech or text. It is to say that the speech community awards the expression a certain social standing, that it is a *social institution*. The nature of the award varies across expressions. A very common kind of status award to an expression (but not the strongest kind) is recognize it as *the name of* or *term for* a class of referents (term) or a unique referent (proper name). The notions 'name' and 'term' have no place in the grammarians' view of lexicon.[8]

Some terms have the full weight of the legal system behind them. One may go to jail if the judge or jury decides that *the weight of the evidence* indicates that

8. At least 27 kinds of social and linguistic markers of conventional status can be distinguished (Pawley 1986). Among the most systematic research on the lexical status of compounds is that done by anthropologists and linguists dealing with folk taxonomies, which systematizes ideas and practices that were already present but often poorly developed in conventional dictionaries.

one's actions can be accurately described as *driving without due care and attention*, or *with intent to injure*, or *with malice aforethought*. One can be acquitted of the charge of uttering a *malicious falsehood* if one's words are judged to be *fair comment*. But the power of legal terms, backed by the trappings of the legal and the judicial systems, is really just a step or two beyond that of ordinary terms, such as *front door* and *back door*, or *apologise* and *ask for permission*, which are deeply embedded in social values and practices. For instance, in English-speaking societies the *front door* and *back door* of a house have different social rank, different appearances and different functions.

How did formulaic expressions fare in humanist treatments of language? Although the literati have always given a bad press to clichés and other stereotyped expressions, the importance of speech formulas in both ordinary language and in specialised genres was recognised, in the decades before the 1970s, by scholars in at least nine different disciplines, besides grammarians.

1. Literary scholars working on epic sung poetry.
2. Anthropologists and folklorists concerned with ritual speech and song and performance routines.
3. Lexicographers.
4. Language teachers and translators.
5. Philosophers concerned with the role of speech acts in ordinary language use and philosophical questions of reference, intention, etc.
6. Sociologists concerned with conversation as strategic interaction.
7. Neurologists and neuro-psychologists, concerned with localisation of language functions in the brain.
8. Psychologists concerned with learning and speech processing.
9. Educational psychologists connecting patterns of language use with patterns of thinking and learning.

By way of example I will refer just to two of these lines of research.

In oral formulaic literary studies the most influential work was that of Milman Parry and Albert Lord on the role of formulas in epic sung poetry (Lord 1960; Parry 1928, 1930, 1932). Parry and Lord recorded in the nick of time the South Slavic tradition of simultaneously singing and composing epic poems in public performances, which still flourished before World War II. Studying the skills of the illiterate Yugoslav singer-composers provided them with a living laboratory in which to test hypotheses about the composition and transmission of Homeric poetry and to demonstrate that this was an oral tradition in which formulas played a central role.

Parry (1930: 80) defined a formula as "a group of words which is regularly employed under the same metrical conditions to express a given essential idea". Parry and Lord recognized that formulas are at the same time both memorized

and flexible, allowing the singer to insert creative variations while maintaining fluency. A 'substitution system' is a group of formulas which show lexical substitutions expressing the same basic structure and idea, or which express the same basic idea with varying number of syllables, enabling the poet to meet a range of different metric conditions.

In English lexicography before the 1970s, formulas were acknowledged in one of two ways. General dictionaries included a selection of phrasal expressions as secondary entries under primary headwords. And a handful of general phrasal dictionaries were compiled, such as Eric Partridge's *Dictionary of Cliches* and *Dictionary of Catch Phrases*, along with more specialised compilations, such as dictionaries of proverbs. Most lexicographical treatments of speech formulas from this era were generally crude and unsophisticated, with information about many of the formal or functional variables either completely missing or given very imprecisely. However, there were notable exceptions, chiefly works on English phraseology for EFL students by H.E. Palmer and A.S. Hornby between the late 1920s and early 1940s, which drew attention to the prevalence of collocations in ordinary language and tackled the syntactic analysis of phrasal expressions, and pioneering work by East European scholars from the late 1940s on. As A.P. Cowie has pointed out (Cowie 1998), an important insight from the East European work was a distinction between several types of phrasal expressions that are often all loosely classed as idioms: those that Cowie calls 'pure idioms' (*kick the bucket*, *bite the dust*, *spill the beans*, *shoot the breeze*) are relatively rare. Two other types are much more numerous: 'figurative idioms' where the words hint at the meaning (*keep s.o. on their toes, run rings around s.o.*, *go off the rails*), and 'restricted collocations' where the base carries a sense that it only has when paired with a collocate (*meet the demand, beg the question, commit suicide, champion a cause, run a deficit, blow a fuse, be sound asleep, chequered career, pitched battle*). Given the large amount of polysemy in common words, the number of restricted collocations is probably much larger than any phrasal dictionary of English has recorded.

The humanist view is incompatible with the grammarian's view in that (1) in the former, the notions 'lexical unit' and 'lexicalised' are usage-based and not grammaticality-based. The lexicon is not a residue of irregular form-meaning pairings but a store of conventional expressions, an (2) more generally, the notion 'language' is broader, resembling Hymes' 'communicative competence'.

4. On some projects from the 1970s investigating formulaic language

4.1 Suddenly formulas are in the air everywhere – but why?

In theoretical linguistics in 1970 the front page story was still transformational-generative grammar but references to speech formulas were creeping into the back

pages. Instead of striving for a monolithic model of language, concerned only with the central systems, some linguists began to pay close attention to the peripheral systems, sometimes encroaching on territory that had previously been mainly of interest to humanists. Grammarians had already pointed out that idioms are a problem for generative syntax (Chafe 1968; Fraser 1970; Makkai 1972; Weinreich 1969). Suspicion was growing that idioms are just the tip of the iceberg and that prefabricated units, including speech formulas, play a much bigger role in ordinary linguistic behaviour than had previously been imagined. Evidence for this emerged from several diverse lines of research, e.g., work on English phrasal lexicography, on discourse and conversation structure, on 1st and 2nd language acquisition, on pragmatics, in work on what was to become frame semantics and construction grammar, on language pedagogy, and on language and the brain, among others. In the beginning, as far as I can tell, these groups of researchers were often unaware of each other's work and had no common theoretical agenda, so one wonders what sparked off this flurry of separate projects.

Some of the intellectual connections are reasonably clear. Some linguists were no doubt stimulated, or provoked, by the strong claims and the hubris of the early years of generative grammar to do work, for example, on language acquisition, idioms and selectional restrictions. On the other hand, the compilation of the first sophisticated phrasal dictionaries of English, Cowie and Mackin's *Oxford Dictionary of Current Idiomatic English. Vol. 1: Verbs with Prepositions and Particles* (1975*)* and its companion, Cowie, Mackin and McCaig's *Vol. 2: English Idioms* (1983) owe something to the discovery by western scholars of Eastern European research in phraseology. Work on speech acts by philosophers of ordinary language stimulated new work in pragmatics by syntacticians that tried to integrate speech act functions into generative syntax via performative verbs. Work on hesitation phenomena and cognition in experimental psychology stimulated research by linguists on speech processing; work on oral epic poetry and in the ethnography of speaking stimulated studies of both oral formulaic genres of discourse and ordinary language. In the UK studies of discourse (e.g., Sinclair & Coulthard 1975), as distinct from grammar, built on Halliday's hierarchy of discourse categories which were stimulated by J.R. Firth's dictum that conversation is the basic form of language use.

The range of work done on formulaic language and related matters in the 1970s and later is too large to review here. For general surveys the reader is referred to Pawley (2007), Wray (2002) and Cowie (ed. 1998) and, for surveys of work on formulaic language and the brain, to van Lancker (1987, 1997). Here I will focus on two projects that are likely to be little known to most theoretical linguists. Both were centred in New Zealand, both began independently but produced parallel findings.

4.2 Kuiper on oral formulaic genres

Perhaps the most impressive body of analytic work on formulaic speech in English is that built up over the past 30 years by Koenraad Kuiper and his associates at the University of Canterbury in Christchurch, New Zealand. In the early 1970s Kuiper, then a young lecturer teaching general linguistics and working in generative syntax, began to visit livestock auctions at Addington in North Canterbury. He was armed with a small, unobtrusive tape recorder and his initial purpose was to record the narrative speech of country folk. He felt this would provide an additional, and perhaps more representative body of data on ordinary spoken English than the upper middle class conversations given in Crystal and Davy's recently published book, *Investigating English Style* (1969). But he couldn't avoid hearing the auctioneers' sales talk, distinctively loud, rhythmic, droned, and full of formulas – and having read Lord's *Singer of Tales* and Beowulf, he immediately recognized that they were oral formulaic performers. What did they have in common with the oral epic poets and why did these commonalities exist?

Kuiper went on to look at other kinds of auctions and several kinds of sports commentary from several countries: Australia, England and the USA, as well as New Zealand, and at a wider range of spoken and written genres, usually working with students or colleagues (e.g., Hickey & Kuiper 2000; Kuiper 1992, 1996; Kuiper & Austin 1990; Kuiper & Haggo 1985; Kuiper & Flindall 2000; Kuiper & Tillis 1986; Flindall 1991; Hickey 1991). General overviews with discussion of theoretical implications are given in Kuiper (1996, 2000).

He found that oral formulaic speech traditions show five features that, taken together, distinguish them from other discourse genres:

a. very strict discourse structure rules, specifying the topics proper to the discourse and their order of occurrence. The discourse structure is hierarchical and can be formally represented by context-free rewrite rules (with a few extra notational conventions). For example, in stock auctions (Kuiper and Haggo 1984) there are four compulsory immediate constituents: (1) *Description of the lot,* (2) *Search for the first bid,* (3) *Calling the bids*, and (4) *Sale*. Most of these constituents in turn may consist of several constituents, e.g., *Description* can consist of *Provenance* + *Number, History, Preparation*, and *Potential*.
b. a very high concentration of speech formulas (usually 90 percent or more of clauses), each anchored to a particular discourse context or range of discourse contexts.
c. special grammatical rules applying to formulas.
d. special prosodic or musical patterns.
e. exceptional fluency, i.e., fewer than average unplanned pauses within clauses.

Although most formulae have the syntactic structure of normal phrases or sentences, in performance they are stored and used as automatic chains. That is to say, once a formula is selected the speaker typically encodes one lexical unit one after another, making a choice among alternatives in slots where there are choices, without take account of higher levels of syntactic structure, as must be done when generating novel sentences of some syntactic structures. In formal terms, individual formulae appear to be generated by finite state grammars (Markov chains) with loops.

Why does the speech of auctioneers and sports commentators have these characteristics? Kuiper and Haggo conclude that the oral formulaic technique has evolved to allow the performer to maintain exceptional fluency while also achieving acceptable standards of content and delivery. They note the close parallels with the Yugoslav oral poets in the need to retain the attention of a mobile audience, in the heavy load placed on the short term memory, in the dense employment of formulae, in the methods by which neophyte practitioners learn their craft and become virtuosos. There are differences: the auctioneer interacts with his audience during the performance. In auctioneers' talk there is less creative imagery. However, it is characteristic of both types that performers do *not* rely on verbatim recall. Perfect recall of long stretches of text requires exceptional concentration and can detract from other facets of performance (note that recall of text by stage actors is a very different task). A more efficient technique is to draw on memorised chunks but to be able to vary the text somewhat and this is what auctioneers and epic chanters do.

Kuiper & Haggo (1984) and Kuiper (1996) outline a descriptive framework for describing oral formulaic discourse. The descriptions are intended to be generative in two senses. First, they seek to be explicit, defining in a precise manner the object of inquiry and its structure. Second, they seek to be predictive, formulating rules for the production of acceptable utterances or texts that go beyond the corpus of recorded examples.

Kuiper's work takes several steps towards achieving the goal that Chafe sets in the quote at the head of this paper. It provides a framework both for describing the structure of a family of discourse genres and for explaining the structure.

4.3 Pawley and Syder on the role of formulas in ordinary language

However, formulaic genres are, plainly, a special class. Kuiper raised the question of how far the characteristics of oral formulaic discourse are unique to that type and how far they are part of ordinary language, e.g., conversation, and spontaneous narrative speech. It happened that I had been reflecting on this particular problem for some time at the University of Auckland. When Kuiper and I met at that inaugural national conference of the Linguistic Society of NZ in 1976 we

were amazed to find that we had been working on similar theoretical problems, provoked by similar experiences, and had arrived at similar answers.

Our academic backgrounds were rather different. I was an anthropological linguist, with some training in experimental psychology and anthropology, who worked on Austronesian and Papuan languages. Ever since I was a struggling student of French at a small town high school in New Zealand I'd been interested in the question of what it takes to achieve expert command of a foreign language. At school I was dimly aware of the disheartening fact that a perfect knowledge of the grammar and of all the words in the lexicon would not come close to making me a fluent and idiomatic speaker of French. There was a lot of other stuff that had to be learned, including thousands of idiomatic ways of saying particular things. My awareness gained a sharp edge between the ages of 17 and 27, when I sought to become a reasonably proficient speaker of half a dozen Pacific Island languages and worked on close analysis of several of these.

Then, mainly owing to the work of my mother, Frances Syder, an English teacher, I also developed an interest in English conversational speech. Between 1972 and 1976 Syder and I collaborated in a project transcribing and analysing a sizeable corpus of English conversational speech recorded in New Zealand and Tasmania. The transcribing work revealed some obvious generalisations about patterns of fluency. I had already read work in experimental psychology on hesitation phenomena and speech processing, on the problem of serial order in behaviour, and on limitations on short-term memory capacity. And, luckily, one of our transcribers, an old schoolmate of mine, had studied Latin and Greek literature and pointed me to the work of Parry and Lord on Homer. Suddenly a number of things fell into place.

The paper we gave at the 1976 LSNZ conference was called 'The one clause-at-a-time hypothesis'. It addressed the puzzle of nativelike fluency – the paradox that in order to speak a language like a native one must produce fluent chunks that contain more information units than the short term memory can hold. There is evidence indicating that, in one speech planning act, speakers cannot encode novel lexical combinations across independent clause boundaries.[9] Speakers overcome this mismatch by 'chunking', i.e., by memorising many multiword units and retrieving them as wholes. Like Kuiper, we were looking at what competence in oral performance entails. One implication of our findings, and Kuiper's, was that if you push performance models to the periphery of linguistic theory, you miss a major source of explanations of why languages are organised the way they are (and, indeed, of why they change they way they do).

9. Later published as Pawley and Syder (2000).

In 1977 we drafted 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency', later published as Pawley and Syder (1983a). Besides addressing the problem of nativelike fluency that paper made the following claims (reiterated in Pawley 1985, 1986):

1. Only a small proportion of grammatical strings are nativelike (idiomatic, in the sense of being how native speakers normally say things). The puzzle of nativelike selection is how speakers know which grammatical strings are nativelike and which are not.
2. Idiomatic (nativelike) command of a language rests to a large degree on knowing thousands of 'lexicalized sentence stems'. These are clause- or multi-clause-sized constructions that contain some slots that are lexically specified and others that are filled by abstract grammatical categories. Today I'd call them productive speech formulas.
3. Each of these (semi) productive formulas has its own mini-grammar. An example is the formula:

> *If it be*-TENSE *good enough for NP (to* S) *it be*-TENSE *good enough for* Y

where in order to justify doing something that others might question as socially unacceptable the speaker refers to the example of an authority figure, as in:

> *If it is good enough for the* ***Queen*** *to wear polka-dot slacks to church it's good enough for me*

Here TENSE (probably) must be either simple present or simple past and the tenses of the two *be* verbs must either agree or be PAST and PRESENT, *respectively.* NP must refer to an appropriate authority figure.
4. However, 'mini-grammar' is not an entirely appropriate description of the constraints on possible lexical substitutions or grammatical expansions in a formula. Breaking these constraints by using normal grammatical options result in utterances that are unidiomatic but not ungrammatical. Thus a distinction must be made between 'grammaticality' and 'idiomaticity' constraints. The distinction can be illustrated by the English time-telling formula:

> *The time/It be*-TENSE M *to/past* H

where M is an expression specifying quantity of time before or after the hour and H specifies the hour. There are severe constraints on how M and H can be expressed, idiomatically. You can say *(The time is*) *ten past five, a quarter past five*, *half past five, twenty to six,* but it is not natural to say *(The time is) a third to six, two thirds past five, five and three quarters, half before six*, six less 20, or *half past five plus ten.* Similar kinds of constraints apply to many others ways of talking about quantity, e.g., height, weight, distances and prices.

5. Lexicalized sentence stems typically have special discourse functions and the constraints on their form are tied to these functions.
6. The model of linguistic competence yielded by this analysis allows no sharp break between grammar and lexicon. Instead there is a continuum of more or less lexicalized constructions, with their own 'grammar', lying between the extremes of abstract constructions and unanalysable lexical units.
7. There are implications for explanatory adequacy. It appears that competent speakers of a language know many linguistic entities in two ways: holistically and analytically, and can move between the two. People are good at generalising, at perceiving patterns, and the generalising capacity is essential to the learning of general rules. On the other hand, people have severely limited rapid processing capacity but they have an enormous memory, which allows them to store and retrieve, or recognize familiar complex form-meaning pairings. Thus, a realistic account of the cognitive processes that underpin nativelike command of a language should accommodate this kind of dual knowledge.

5. Have linguists changed their views of language?

Let me jump forward to 2007. Usage-based or performance-based approaches have now figured quite prominently in linguistic research for more than 30 years, producing a vast body of data and analysis on ordinary language use. The very term 'usage-based' implies common ground with the approach of dictionary-makers but it is a term that covers a range of approaches by scholars with diverse theoretical agendas (Barlow and Kemmer 2000). In this final section I want to ask whether paying close attention to usage and discourse has led many linguists to develop models of what languages consist of that differ from those that were prevalent among grammarians in 1970; or indeed, from those that emerged during the 1970s, in the early years of usage-based work.

There is no doubt that we have learnt much more about the so-called 'peripheral' parts of languages. It has turned out that the periphery is much important than was once thought. We know more about the types of minor constructions and conventional expressions and that such entities are prevalent in both spontaneous speech and written discourse. As Cowie has observed, whereas in the early 1980s "it was still possible to dismiss phraseology as a linguistic activity of only minority interest and with poor prospects of recognition as a level of language or of linguistic description" except in dictionary-making (Cowie 1998: 18), it "has now become [a] major field of pure and applied research for Western linguists" (Cowie 1998: 1).

Have some radical proposals about how languages work and how they are acquired come out of discourse-based linguistics? Yes, to some degree. One thinks of,

among others, of Chafe (1979, 1980, 1994) and others on how thinking and speaking are connected in the encoding process, of Givón (1979) on grammar as a processing mechanism, of Hopper (1987) on emergent grammar and of Grace (1981, 1987) on 'saying things' and the linguistic construction of worlds. Under the rubric of construction grammar (Fillmore et al. 1988; Croft 2001; Goldberg 1995, 2006; Tomasello 2003 and Wray (2002) and cognitive grammar (Langacker 1987, 1991) there have been a range of proposals that question the need to posit a boundary between lexicon and grammar – so-called 'continuum models' that regard all the work of pairing form and meaning as falling to constructions. Some of these proposals have much in common with those of Kuiper and Pawley and Syder reviewed above.

Nevertheless, certain methodological and theoretical biases have, in my view, acted as conservative forces, helping to preserve traditional ways of viewing language, in particular, continuing to focussing on the form of sentences and phrases much more strongly than on their communicative and cognitive functions.

The creation of large electronic corpora and efficient search engines has been a great help in studying the frequencies of collocations and construction types but this powerful machinery has encouraged linguists to produce a lop-sided account of what is important for language learners. Frequency of use is an important part of linguistic experience but it is only one part. There is also the social, contextual and dramaturgical baggage associated with acts of speaking. There are payoffs, consequences. One gets rewarded or otherwise for saying and doing certain things. The language learner finds that certain expressions are associated with certain gestures and voice qualities, as well as with certain physical and social contexts, purposes and consequences. Conversational speech and various other forms of discourse seem to be highly structured in terms of the norms of what things can be said, and when, why and how. It is clear that minor constructions and speech formulas play an absolutely central role in everyday spoken discourse, contributing to its fluency, idiomaticity, coherence, appropriateness, and wit, but some of these qualities are not easily located and counted. Machine searches need to be supplemented with qualitative analysis but of course this is harder to do.

By and large, it seems to me that most linguists doing discourse-based research still seek to preserve the old grammar-lexicon model by extending the terms 'grammar', 'grammaticality' and 'grammaticalization' to accommodate new phenomena. It is no accident that the prestige term 'grammar' keeps popping up in the names of new usage-based approaches to language: 'construction grammar', 'cognitive grammar', 'space grammar', 'emergent grammar', 'pattern grammar', etc.

Construction grammar has advocated a rethinking of the cognitive basis of linguistic competence based on what is entailed in learning minor constructions (productive speech formulas). Thus, Goldberg (2006: 14) writes that

> … all linguists recognize that a wide range of semi-idiomatic constructions exist in every language, constructions that cannot be accounted for by general, universal or innate principles or constraints.… Generative linguists argue that these constructions exist only on the "periphery" or "residue" of language – that they need not be the focus of linguistic or learning theorists. Constructionists on the other hand [argue] that whatever means we use to learn these patterns can easily be extended to account for so-called "core" phenomena. In fact, by definition, the core phenomena are more regular, and tend to occur more frequently … Therefore, if anything they will be easier to learn.

But are these 'semi-idiomatic constructions' part of grammar, or lexicon, or something else?

In their paper on the 'let alone' construction Fillmore, Kay and O'Connor (1988: 534) conclude that

> *in the construction of a grammar* more is needed than a system of general grammatical rules and a lexicon of fixed words and phrases … [A] large part of a language user's competence is to be described as a repertory of clusters of information including, simultaneously, morphosyntactic patterns, semantic interpretation principles to which these are dedicated … and in many cases, specific pragmatic functions in whose service they exist. … (my italics: AP)

Here Fillmore *et al.* advocate dispensing with the old idea that languages consist of a grammar and a lexicon. But they wish to extend the term 'grammar' and by implication, 'grammaticality', to cover various things that did not used to be subsumed under this rubric. I am guilty of this, too. Pawley and Syder (1983a: 216), discussing how to handle productive speech formulas (lexicalized sentence stems), write that "Each dictionary entry for [such an entity] will, presumably, be a mini-grammar" and go on to specify various kinds of information that will be in the entry including idiomaticity constraints and functions. It is hard to escape deeply-ingrained ways of talking.

Chafe (1996:459) seeks "a model of natural discourse that pulls together the diverse cognitive and social factors responsible for the shape of language". Many of these diverse factors can be seen at work in speech formulas. To understand the part played in command of English by typical speech formulas, such as *a stitch in time saves nine* or *If it's good enough for X (to S), it's good enough for me*, or *Would you care/like to join us*?, we need to investigate not only their grammatical and lexical makeup, semantic and pragmatic meanings, intonation patterns, voice quality, etc. but also their communicative functions, their role in constructing discourse that is coherent, socially appropriate, strategically effective, poetic, witty, etc. and their role in speech processing, e.g. as prefabricated schemas that underpin fluent and idiomatic speech.

At the level of the nitty-gritty, it seems to me that studies by construction grammarians of semi-idiomatic constructions (speech formulas) have been largely pre-

occupied with grammatical and semantic structure. Other components of such constructions have yet to receive due attention – for example, their social and discourse functions and their musical (prosody, voice quality, etc.) and body language components – and their roles in speech production and comprehension and language learning (there are exceptions, e.g. Tomasello 2003; Wray 2000, 2002). That is natural: analysis of grammar is what grammarians have always done best. Some impressive formalisms have been developed (or borrowed) to describe the grammar and semantics of conventional expressions (e.g. Fillmore et al. 1988; Kay & Fillmore 1999; Kuiper 2000) but apparatus of comparable sophistication for handling these other elements is still lacking. There are plenty of challenges ahead of us.

References

Austin, John Langshaw. 1962. *How to do things with words.* Oxford: Clarendon Press.

Barlow, Michael & Suzanne Kemmer. 2000. *Usage-based models of language.* Stanford CA: CSLI.

Bloomfield, Leonard. 1933. *Language*. London: Allen & Unwin.

Chafe, Wallace. 1968. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language* 4: 109–127.

Chafe, Wallace. 1979. The flow of thought and the flow of language. In *Syntax and semantics*. Vol. 12: *Discourse and syntax*, 159–181. New York NY: Academic Press.

Chafe, Wallace. 1980. The development of consciousness in the production of a narrative. In *The pear stories. Cognitive, cultural and linguistic aspects of narrative production*, W. Chafe (Ed.), 9–50. Norwood NJ: Ablex.

Chafe, Wallace. 1994. *Discourse, consciousness and time.* Chicago IL: University of Chicago Press.

Chafe, Wallace. 1996. Beyond beads on a string and branches in a tree. In Goldberg, 49–65.

Chomsky, Noam. 1964. The logical basis of linguistic theory. In *Proceedings of the Ninth International Congress of Linguistics*, H. Lunt (Ed.), 914–78. The Hague: Mouton.

Chomsky, Noam. 1965 *Aspects of the theory of syntax*. Cambridge MA: The MIT Press.

Chomsky, Noam. 1970. Remarks on nominalizations. In *Readings in English transformational grammar*, R. Jacobs & P. Rosenbaum (Eds), 194–221. Waltham MA: Ginn & Co.

Cowie, Anthony P. 1998. Phraseological dictionaries: Some East-West comparisons. In *Phraseology: Theory, analysis and application,* A.P. Cowie (Ed.), 209–228. Oxford: Clarendon Press.

Cowie, Anthony P. & Ronald Mackin 1975. *Oxford dictionary of current idiomatic English,* Vol. 1: *Verbs with prepositions and particles.* Oxford: OUP.

Cowie, Anthony P., Ronald Mackin & Isabel R. McCaig 1983. *Oxford dictionary of current idiomatic English,* Vol. 2: *English idioms.* Oxford: OUP.

Croft, William, 2001. *Radical construction grammar. Syntactic theory in typological perspective*. Oxford: OUP.

Crystal, David & Derek Davy. 1969. *Investigating English style*. Cambridge: CUP.

Fillmore, Charles J., Paul Kay & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64: 501–538.

Flindall, Marie. 1991. Checkout operators: Formulae and oral traditions. Ms, University of Canterbury, New Zealand.

Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of Language* 6: 22–42.

Givón, Talmy. 1979. *On understanding grammar.* New York NY: Academic Press.
Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago IL: University of Chicago Press.
Goldberg, Adele E. 2006. *Constructions at work. The nature of generalization in language*. Oxford: OUP.
Goldberg, Adele E. (Ed.), 1996. *Conceptual structure, discourse and language.* Stanford CA: CSLI.
Grace, George W. 1981. *An essay on language.* Columbia SC: Hornbeam Press.
Grace, George W. 1987. *The linguistic construction of reality.* New York NY: Croom Helm.
Hickey, Francesca. 1991. What Penelope said: Styling the weather forecast. MA thesis, University of Canterbury, New Zealand.
Hickey, Francesca & Koenraad Kuiper. 2000. A deep depression covers the South Tasman Sea; New Zealand Met Office weather forecasts. In *New Zealand English*, A. Bell & K. Kuiper (Eds), 279–296. Wellington & Amsterdam: VUW Press & John Benjamins.
Hockett, Charles F. 1958. *A course in modern linguistics.* New York NY: MacMillan.
Hopper, Paul. 1987. Emergent grammar. *Berkeley Linguistic Society* 13: 139–157.
Hymes, Dell. 1962. The ethnography of speaking. In *Anthropology and human behavior*, T. Gladwin & W.C. Sturtevant (Eds), 13–53. Washington DC: Anthropological Society of Washington.
Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations. *Language* 75(1): 1–33.
Kuiper, Koenraad. 1985. The nature of ice hockey commentaries. In *Regionalism and national identity: Multidisciplinary essays on Canada, Australia and New Zealand*, R. Barry & J. Acheson (Eds), 167–175. Christchurch: Assoc. for Canadian Studies in Australia and New Zealand.
Kuiper, Koenraad. 1992. The English oral tradition in auction speech. *American Speech* 67: 279–289.
Kuiper, Koenraad. 1996. *Smooth talkers. The linguistic performance of auctionners and sportscasters.* Mahwah NJ: Lawrence Erlbaum Associates.
Kuiper, Koenraad. 2000. On the linguistic properties of formulaic speech. *Oral Tradition* 15(2): 279–305.
Kuiper, Koenraad & Paddy Austin. 1990. They're off and racing now: The speech of the New Zealander race caller. In *New Zealand ways of speaking English,* A. Bell & J. Holmes (Eds), 195–220. Clevedon: Multilingual Matters.
Kuiper, Koenraad & Marie Flindall. 2000. Social rituals, formulaic speech and small talk at the supermarket checkout. In *Small Talk*, J. Coupland (Ed.), 183–207. London: Longman.
Kuiper, Koenraad & Douglas Haggo. 1984. Livestock auctions, oral poetry and ordinary language. *Language in Society* 13: 205–234.
Kuiper, Koenraad & Frederick Tillis. 1986. The chant of the tobacco auctioneer. *American Speech* 60: 141–149.
Langacker, Ronald. 1987. *Foundations of cognitive grammar,* Vol. 1. *Theoretical prerequisites*. Stanford CA: Stanford University Press.
Langacker, Ronald. 1991. *Foundations of cognitive grammar.* Vol. 2. *Descriptive applications.* Stanford CA: Stanford University Press.
Lord, Albert. 1960. *The singer of tales.* Cambridge MA: Harvard University Press.
Lyons, John. 1968. *Introduction to theoretical linguistics.* Cambridge: CUP.
Makkai, Adam. 1972. *Idiom structure in English*. The Hague: Moutom.
Newmeyer, Fredrerick. 1986. *The politics of linguistics*. Chicago IL: University of Chicago Press.
Palmer, Harold Edward. 1933. *Second interim report on English collocations.* Tokyo: Kaitakusha.
Palmer, Harold Edward. 1938. *A grammar of English words.* London: Longmans Green.
Parry, Milman. 1928. *L'Epithete traditionelle dans Homère.* Paris.

Parry, Milman. 1930. Studies in the epic technique of oral verse-making. I: Homer and Homeric style. *Harvard Studies in Classical Philology* 41: 73–147.

Parry, Milman. 1932. Studies in the epic technique of oral verse-making. II: The Homeric language as the language of an oral poetry. *Harvard Studies in Classical Philology* 43: 1–50.

Pawley, Andrew. 1985. On speech formulas and linguistic competence. *Lenguas Modernas* (Chile) 12: 84–104.

Pawley, Andrew. 1986. Lexicalization. In *Georgetown Round Table in Languages and Linguistics 1985. Languages and linguistics: The interdependence of theory, data, and application,* D. Tannen & J. Alatis (Eds), 98–120 . Washington DC: Georgetown University.

Pawley, Andrew. 2007. Developments in the study of formulaic language since 1970: A personal view. In *Phraseology and culture in English*, P. Skandera (Ed.), 3–34. Berlin: Mouton de Gruyter.

Pawley, Andrew & Frances Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication*, J.C. Richards & R.W. Schmidt (Eds), 191–227. London: Longman.

Pawley, Andrew & Frances Syder. 2000. The one clause at a time hypothesis. In *Perspectives on fluency*, H. Riggenbach (Ed.), 167–191. Ann Arbor MI: University of Michigan Press.

Peters, Ann. 1983. *The units of language acquisition*. Cambridge: CUP.

Quirk, Randolph, Geoffrey Leech & Jan Svartvik. 1972. *A grammar of contemporary English.* Cambridge: CUP.

Robbins, Robert H. 1967. *A history of linguistics*. London: Longman.

Saussure, Ferdinand de. 1916. *Course in general linguistics*. Geneva: University of Geneva.

Schrempp, Gregory. 1992. *Magical arrows. The Maoris, the Greeks, and the folklore of the universe.* Madison WI: University of Wisconsin Press.

Searle, John. 1969. *Speech acts: An essay in the philosophy of language.* Cambridge: CUP.

Seuren, Peter. 1969. *Operators and nucleus. A contribution to the theory of grammar.* Cambridge: CUP.

Sinclair, John & Malcolm Coulthard, 1975 . *Towards an analysis of discourse.* London: OUP.

Tomasello, Michael. 2003. *Constructing a language. A usage-based theory of language acquisition.* Cambridge MA: Harvard University Press.

van Lancker, Diana. 1987. Non-propositional speech: Neurolinguistic studies. In *Progress in the psychology of language*, Vol. 3, A.W. Ellis (Ed.), 49–118. Hillsdale NJ: Lawrence Erlbaum Associates.

van Lancker, Diana. 1997. Rags to riches: Our increasing appreciation of cognitive and communicative abilities of the human right cerebral hemisphere. *Brain and Language* 57: 1–11.

Weinreich, Uriel. 1969. Problems in the analysis of idioms. In *Substance and structure of language,* Jaan Puhvel (Ed.), 23–81 . Berkeley CA: University of California Press.

Wray, Alison. 2000. The functions of formulaic language: An integrated model. *Language and Communication* 20(1): 1–28.

Wray, Alison. 2002. *Formulaic language and the lexicon.* Cambridge: CUP.

Identifying formulaic language

Persistent challenges and new opportunities

Alison Wray
Cardiff University, **UK**

Abstract

Identifying examples of formulaic language in text is a non-trivial challenge, but the difficulties can be much alleviated by the use of an appropriate definition. Three types of definition are distinguished. Type (i) lays out an analytic working space. Type (ii) derives from an analysis and represents a theoretical position. Type (iii) locates examples for subsequent analysis. Examples of each type are discussed.

Extreme examples of formulaicity (pre-memorized material, political slogans and military bugle calls) are then used to explore the boundaries of the definition of formulaicity as **morpheme-equivalence**. Addressing the question 'Do formulaic sequences constrain expression?' reveals the inherent tension between novelty and formulaicity in balancing processing parsimony and the need to respond appropriately in unique communicative events.

1. Introduction

Researching formulaic language has many challenges, but probably the single most persistent and unsettling one is knowing whether or not you have identified all and only the right material in your analyses. In order to answer interesting questions about the nature of formulaicity, and to explore and challenge the claims and predictions of theoretical models, it is important to have reliable examples of the phenomenon under scrutiny. The appropriate identification of examples depends on the definition used, and different types of definition are appropriate to different purposes. In the first half of this chapter I shall offer an explanation for why there are difficulties with identification, and explore ways of ensuring that the right choice of definition is made.

For most researchers, the nub of the problem with identification is figuring out where novel language stops and formulaic language begins. This is because the dynamics of effective communication can be argued largely to occur at that boundary. However, in the second half of this chapter I shall argue that one can never fully define a phenomenon unless one has walked **all** of its boundaries. To demonstrate the point, I shall explore what happens at the **least** novel end of formulaicity in communication. Certain theoretical claims about formulaicity can be investigated quite effectively by looking at such extreme examples, because the opportunities for 'escaping' into novel expression are considerably reduced. That makes it is easier to observe the underlying forces that determine how and when the transition between novelty and formulaicity is made. I shall ask what happens in certain situations in which communication is obliged, to a greater or lesser extent, to remain formulaic, even when novel expression is desirable.

2. Approaches to definition

The challenge of defining a phenomenon begins with how to **talk** about defining it. How is one to refer to that phenomenon independently of potential definitions? Overall, researchers should be wary of vagueness in terms, but sometimes the only way to proceed is to designate one or two terms to remain deliberately vague. In order to make this chapter viable, the terms 'formulaic language' and 'formulaicity' are used

to refer **generally** to the kind of linguistic material under discussion, without any attempt to delineate exactly where that material starts and stops. By using these terms in this deliberately fuzzy way, it becomes much easier to talk with precision about how other terms, according to their definition, divide up the conceptual space.

2.1 Types of definition

We can identify, for present purposes, three types of definition, each with its own function:

i. definitions you start with, in order to explore the fundamental nature of the defined phenomenon;
ii. definitions you end up with, that describe and explain the defined phenomenon;
iii. definitions you work with, that reliably identify examples so that other questions can be asked about them.

Types (i) and (iii) are both, technically, **stipulative** definitions – that is, the researcher decides what will fall inside and outside the boundaries of the definition, as the basis upon which the analysis proceeds. Type (ii) definitions are **descriptive** – the evidence determines how they are configured, and there are particular opportunities for research in consequence of them. As a route into exploring the three types of definition, let us consider, first, a term that has gained some currency in the recent literature, 'formulaic sequence'. The formulaic sequence was defined by Wray (2002a) as follows:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (p. 9).

It was coined as a means of referring to all the material that might turn out to be formulaic in the ensuing analysis reported in Wray (2002a). It enabled the laying out of a working space, in advance of a detailed consideration of what was in that space. For this reason, it is a type (i) definition.

A type (ii) definition emerges from an analysis, on the basis of evidence, or as a natural product of the theory that is developed. In the course of Wray (2002a: 265–269) a definition develops, according to which formulaicity is described as 'morpheme-equivalence' – explained in more detail below. Definitions of this kind are used to explain patterns in existing data and to make predictions about future data. Importantly, they can also be used as an anchor for testing the theory from which they derive.

A type (iii) definition is the kind you work with **as** you analyse data. It lays out the parameters for identifying examples. Ideally, it will be based on a balance of two considerations: expediency and theoretical plausibility. For instance, it is, for many

researchers, much easier to identify wordstrings that are continuous and frequent than ones that are discontinuous and rare, and this might encourage them to define formulaic language in that way. However, such a definition needs to be justified by a theory: some theories require equal consideration to be given to discontinuous and infrequent items. Conversely, one could take a strong theoretical starting point and then get into practical difficulties with identification. One might hypothesise that an interesting subset of formulaic language is items that the speaker or hearer has heard before (and so has had a chance to store in memory). The problem would be how one can always reliably know what a person has heard before[1].

If we attempt to co-opt the type (i) definition of the 'formulaic sequence' for a type (iii) function, problems arise. The definition describes "a sequence … which is, or appears to be … stored and retrieved whole from memory at the time of use". One cannot go to data and reliably pick out items on that basis: we do not have an independent way to establish that something is "stored and retrieved whole"[2], and there is intrinsic subjectivity in the notion of something "appear[ing] to be" prefabricated. In other words, the features of this definition that make it useful for laying out a working space (type i), make it unfit as a type (iii) definition. For reliably identifying examples in data, then, specific types of definition, dedicated to that purpose, are required. Examples will be reviewed below.

Before we proceed, however, the type (ii) definition 'morpheme-equivalence' needs to be explained, since it will underpin much of what is said in the rest of the chapter.

2.2 Morpheme-equivalence and the blurring of the boundary between formulaic and non-formulaic material

The definition of formulaicity as 'morpheme-equivalence' (Wray 2002a: 265–9) is based on the proposal that certain wordstrings (and also many polymorphemic words) take on characteristics associated with formulaicity (including fluency of

1. Artificial situations can be used to investigate such questions. Wray (2004) observed a Welsh language learner from the very start, so that it was possible to state with considerable confidence what she had encountered before, when and how often, and to use that as an anchor for explaining her language performance. However, such approaches to research are particular and manipulative, and inappropriate for answering many more general questions about the learning and knowledge of formulaic language.

2. Reaction times, reading speed and eye movement are amongst the experimental approaches taken to ascertain whether formulaic sequences are processed more quickly than comparable non-formulaic wordstrings, and with some success (see for instance Conklin & Schmitt 2008). However, it is not possible by such means to differentiate between holistic access and fast-route componential decoding.

production, semantic and/or grammatical oddity, characteristic intonational contours, frequency of occurrence in text) because they have a dedicated entry in the mental lexicon. In line with most models, Wray's takes it that entries in the mental lexicon are atomic, that is, base units of meaning or function that cannot be broken down further. The classic example of such a base unit is the morpheme, the smallest unit of meaning. However, by identifying some internally complex units as morpheme-equivalent, Wray takes the theoretical position that it is possible for lexical entries to contain semantically-viable subparts that are not taken into account during storage and retrieval. In effect, the morpheme-equivalence definition makes a specific prediction: certain kinds of word and wordstring that appear, from their surface form, to be subject to variation, actually will not vary (or, more accurately, will **normally** not vary though they can by special intervention) because their internal composition is not active.

A highly significant feature of the morpheme-equivalence definition, and the theory underpinning it, is that internally complex words and wordstrings become stored as single items **not** (as in many theories) **because of** oddity in their form or meaning, but for a quite unrelated reason: patterns of usage. Therefore, an item **first** becomes formulaic and only subsequently may, as a consequence of being so, accrue certain grammatical, semantic or phonological characteristics typically associated with formulaic language. Items that are formulaic but that have not yet accrued any such features will be indistinguishable from novel configurations, explaining why it can be difficult to identify formulaic language in real text. Their regular formal characteristics will make them resemble something they are not (compositional structures), and not resemble what they are (morpheme-equivalents). One might liken the situation to looking at a completed jigsaw puzzle (Figure 1a), in which

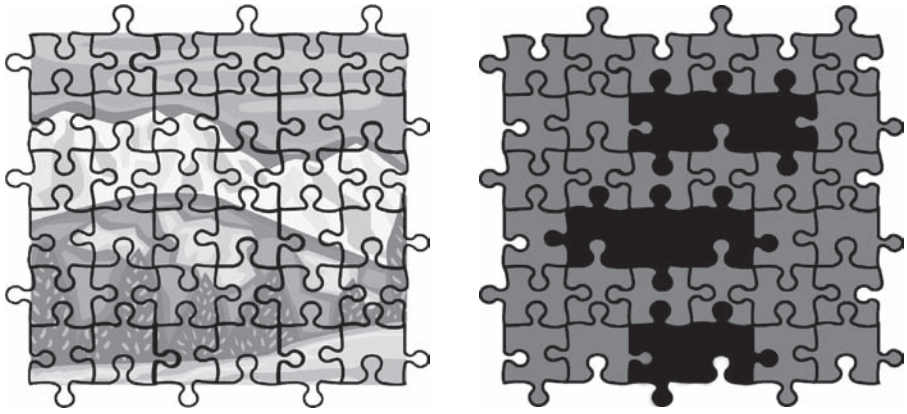


Figure 1(a). Jigsaw puzzle, front view suggests that the picture is constructed from small units. **(b).** Jigsaw puzzle, back view shows that the picture is constructed from both small and large units.

it is possible to see the lines around each individual piece, leading to the assumption that each piece can be moved independently of the others. What is not clear until one looks at the back view (Figure 1b) is that the machine that cut the pieces failed fully to sever the connections, so that there are, in various places, multi-piece strings still joined together.

This suggestion that formulaicity entails the **retention** of inter-word links rather than the **establishment** of them is another distinguishing feature of the theory. The process by which formulaicity is proposed to arise in the individual (as distinct from in the language – see later in this section) is 'needs only analysis' (Wray 2002a: 130–2), whereby one typically begins with multiword strings and breaks them down, rather than with morphemes and words and building up. During L1 acquisition and subsequently, the native speaker[3] attempts to map the largest possible form onto a reliable meaning. The effect is that sometimes words will go around in groups, never separated, because there has been no reason to attribute a sub-part of meaning to a sub-part of form. For example, the three word expression *in order to* will be a holistically stored single lexical unit, because its meaning and function map onto the form as it stands. It is irrelevant to the case that the items *in, order* and *to* also exist in the lexicon as free units because there is nothing about these individual entries that can elucidate the form-meaning relationship of the phrase.[4] Under needs only analysis, most formulaic items in the lexicon will be partly-lexicalized frames, in which there are gaps between fixed parts for the insertion of variable material, including both word endings and open class items. Frames develop when the individual encounters variation within a recurrent pattern of words, and isolates the loci of variation, while keeping the remainder fixed (Peters 1983).

According to needs only analysis, native speakers build up a lexicon of morphemes, words and multiword strings, directly reflecting their experience of patterns of variation in the input. Because of large-unit mapping, people will naturally adopt the turns of phrase typical of their speech community, and develop a sensitivity to what 'sounds right' that is somewhat independent of (i.e., mostly narrower than) the predictions of the grammatical rules of the language.

Finally, we can note how this model explains the tendency for formulaic language to have certain characteristics of form, meaning and/or phonology. A wordstring that is formulaic will be easy and desirable to select and use, since it requires less processing than a novel string of equivalent size. Word bundles are

3. Post-childhood L2 learners are hypothesized also to adopt needs only analysis, but with different outcomes (Wray 2002a, chaps 10 & 11).

4. The fact that etymologists might be able to explain why *in order to* has this form is not relevant to the knowledge of the average language user.

passed from person to person in just the same way that words are. Words can lose their phonological precision and morphological immediacy through holistic form-meaning mapping, until they are etymological relics (placenames such as *Spittalfields* and *Newtown* are a classic example). In the same way, wordstrings, attached to agreed meanings and functions, will adopt new connotations and associations, and, because there is no introspection, may become phonologically distinctive and be protected from grammatical and other changes that the language undergoes over time. The longer an item is insulated in this way, the more it will stand out from the novel material, until it is viewed as irregular in form and/or opaque in meaning (Wray 2002a: 267). Figure 2 represents this process. As an illustration, consider the expression *believe you me!* which in an earlier version of English could be generated by the standard rules of English. We may infer that, becoming a convenient and reliable encoding of the idea 'I'm certain', the wordstring became holistic for many speakers, and subsequently, by virtue of being holistic, did not become modified when the form of the imperative changed in English. In isolation from the active rules of the language, it became impossible to generate as a novel string, and since no novel strings would fall in paradigm with it, it was increasingly unlikely to be broken down. A marker of its formulaicity is its distinctive phonology, be 'lieve 'you 'me. Since the individual words of the expression continue to provide a reliable index of its basic meaning, even if not all of its pragmatic weight, we might locate this wordstring in the 'formulaic, semi-regular' zone of Figure 2.

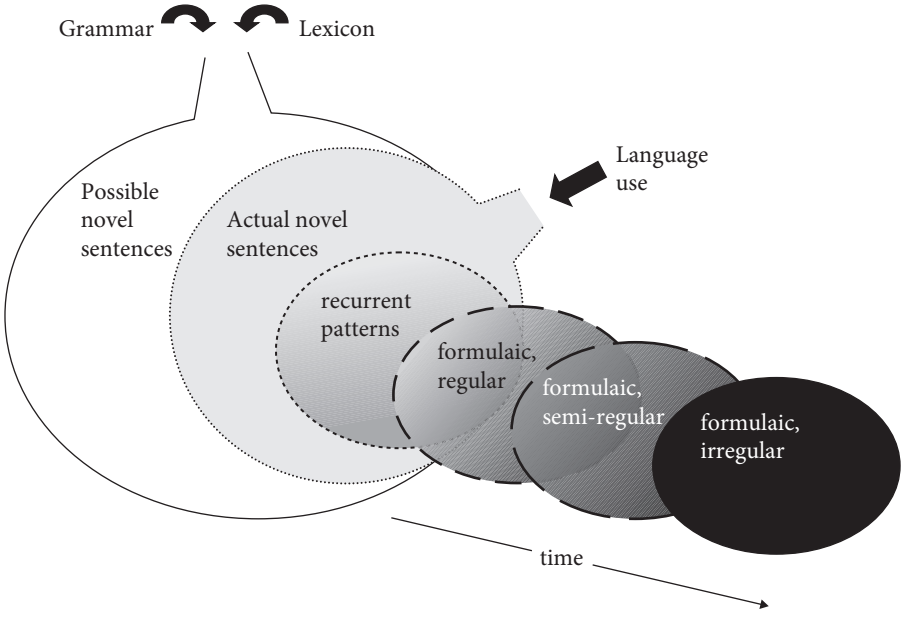


Figure 2. Emergence of increasingly irregular formulaic language from novel language.

Many researchers choose to **define** formulaic language as only those items in the right-most zone (ie. with irregular features of semantics or grammar), an approach that is perfectly valid, provided there is a theoretical account for why formulaic language should consist of all and only such items. The common explanation, that items **first** become irregular and become formulaic as a result, requires a narrative for what motivates irregularity.

What should be clear from this account is that identification, definition and theory are intimately linked, and cannot be pursued other than collectively – identification must proceed on the basis of a sound definition, and definitions must be grounded in theory.

2.3 Harnessing definitions appropriately in research

It is important to acknowledge the purpose of a definition, and select one that can achieve that purpose effectively. Researchers must often eschew others' stipulative (type i and type iii) definitions in favour of producing their own, unless their own purposes substantially overlap with those of the originator. On the other hand a researcher might reasonably adopt another's descriptive (type ii) definition as a focus for examining and interpreting data, and might subsequently use the new evidence to challenge the robustness of that definition and the theory underlying it. For many researchers, it is an effective type (iii) definition that they most urgently need: one able to assist them in talking with confidence about characteristics of formulaicity, using examples from their own data. Therefore, it is this type that will be the focus of the following discussion.

Type (iii) definitions aimed at identifying formulaic language are used for two key identificational purposes. In some research, the aim of identification is to find examples in one or more pieces of data – such as speech or writing by native or non-native speakers – so that the features of those examples can be explored. The most informative analyses will be those that include all appropriate cases, including those on the borderline, so the definition needs to be equipped for decisions about what lies inside and outside its scope. However, other research requires reliable examples **before** the research can be undertaken, because they will be part of the treatment, or input. For instance, Underwood, Schmitt & Galpin (2004) wanted to track readers' eye movements when reading formulaic and non-formulaic material, to see if there was evidence of differences in processing. For that research, it was imperative to identify sets of wordstrings that were sure to be, and not be, formulaic, so borderline cases were best left aside.

We shall not consider the pre-selection cases in much detail, since simply avoiding borderline cases generally makes things straightforward. Nevertheless, it is useful to note a few key issues. Clearly, given the general difficulties with

definition and identification, even selecting input material must be done with care. Making a wrong decision at the design stage could lead to the collection of invalid data. There are various ways in which a researcher can minimize the risks. One is to hook the decisions onto some external, published, justification. In this way, should poor selections turn out to have undermined the investigation, it will be possible both to defend the decisions made, and to challenge the legitimacy of the published claims that underlay them. This is a stronger position to be in than having to admit that one's own intuitive choices – perhaps not strongly grounded in theory – were ill-founded. Other solutions are to use an existing database, or to choose, justify, and stick to, one's own stipulative definition.

Databases are an easy option for those who can legitimately use idioms as input, for there are plenty of lists to choose from. However, beyond idioms there are more potential difficulties, for any list will be predicated on someone's decision about what counts and does not count as formulaic. By adopting a list, one is simultaneously adopting the theoretical assumptions underlying it, embracing any constraints arising from its original purpose (e.g., as a reference resource for foreign language learners), and incorporating any weaknesses in its construction. As Jones & Haywood (2004: 274) observe, little is gained if one rejects one's own intuitions as too subjective for use in research, but then, through the use of a list, adopts someone else's intuitive choices.

Stipulating one's own definition has major advantages, but needs to be carefully justified (see section 2.4). As mentioned earlier, Underwood et al. (2004) required formulaic material for use in their investigation of eye movement. Amongst the features they stipulated were that the wordstrings should have an obvious beginning and end, because it was important to avoid any ambiguity about where the reader first realized the item to be formulaic. Clearly, however, those criteria would have less significance in some other types of study.

2.4 Finding examples of formulaic language in text

Most researchers who have tried to identify all and only the formulaic language in a text have found that there is considerable scope for discussion about what should and should not be counted, and if more than one judge is used, arbitration is going to be required. Although automatic computer-identification is possible, it is a mixed blessing (see 2.4.1). Researchers may, particularly if there is a qualitative focus to their work, need to identify their examples by eye. As already illustrated, definitions will take on the flavour of the researcher's interests and biases – something that research on formulaic language particularly invites, since there are so many different potential criteria to foreground (see Wray; 2002a: chapters 2 & 3). Here

we shall examine just a selection, as a means of asking how definition interacts with identification and theory.

2.4.1 *Can you identify formulaic sequences by counting them?*

For some researchers, the wordstrings that occur most often are the most interesting. There can be very good grounds for that view. For instance, if you are primarily concerned with helping learners work out which wordstrings to spend time becoming familiar with, there is some sense in targeting the ones that turn up in texts most frequently (Ellis, Simpson-Vlach & Maynard 2006).

However, the fact that computer programs can identify the most frequent wordstrings sometimes means that the cart pulls the horse: frequency searches can become the basis for identification purely because they are relatively easy to carry out. Selecting just the most frequent wordstrings may still not be a problem: one always needs some means of sampling, and frequency is one that can be used – others might be choosing the top entry on each page of a dictionary of idioms and clichés, or analysing only wordstrings that begin with the letter S. In the latter two instances, one would be unlikely to believe that the sampling resulted in an unrepresentative set. However, frequency is a less straightforward case, since it may itself contribute to other properties. Perhaps wordstrings can be formulaic whether they are frequent or not, but the more frequent they are, the more likely they are to be irregular in form, or opaque in meaning, or phonologically streamlined (see later sections for consideration of these features). If so, only looking at the frequent examples could distort one's understanding of the wider range of manifestations that formulaic language can have.

Having decided that frequency is an appropriate approach to identification, new challenges arise. Wray (2002a: 28), reviewing approaches to the automatic identification of wordstrings by frequency, notes how the parameters of the definition grossly alter estimates of the amount of formulaic material in the language. Claims made on the basis of data gathered using a particular definition need to be framed within an acknowledgement of how that definition has led to selective identification. For example, one might stipulate the nature of the sequence (e.g., only words adjacent to each other), the length of string (e.g., only sequences of three or more words) and the frequency threshold (e.g., only occurring at least four times per million words). All such stipulations could lead to the exclusion of otherwise relevant items.

Furthermore, the nature of the search could prevent the identification of the full range of manifestations that a single formulaic sequence can naturally take. For instance, will one's definition permit the identification not only of *bite the bullet* but also *bitten the bullet* (ie. can *bite* and *bitten* be treated as instances of the same word in the search?), *bitten this particular bullet* (where a word intervenes

and *the* has been replaced by another determiner), *many such bullets have, over the years, been bitten* (where the word order has changed, other material intervenes, and *bullet* has been pluralized), and so on? There are effective ways to avoid such difficulties if you know they are there; but every analysis is potentially vulnerable to any difficulties not foreseen.

2.4.2 *Can you hear that something is formulaic?*

Van Lancker, Canter & Terbeek (1981) found that it was possible to differentiate, in spoken performance, between the literal and non-literal meanings of idioms. Characteristics of the idiom reading included:

- the key lexical words were shorter in duration than in the literal reading,
- the strings were produced with fewer pauses,
- the pitch contours were less marked,
- the vowels and consonants were less precisely enunciated.

This finding is of considerable use to researchers interested in the less obvious types of formulaic sequence, since it suggests a means of establishing where the boundary lies between formulaic and novel material. Does an expression such as *see you later* or *watch where you're going* have the same phonological characteristics as the literal or non-literal reading of an idiom? A potential complication, however, lies in how one establishes the baseline for comparison – how a reader would produce the **non**-formulaic reading. Van Lancker et al found the pronunciation contrasts only when they asked the speakers to make clear which meaning they intended. Since the items were idioms, it was easy to conceptualise those differences as a non-literal versus literal reading. It would be more difficult to ask speakers to differentiate formulaic and non-formulaic readings of expressions like *see you later*.

2.4.3 *Are formulaic sequences non-canonical?*

Idioms are generally considered formulaic **because** they are non-literal in meaning: one is more or less obliged to treat them holistically, in order to avoid conveying, or interpreting, an inappropriate meaning. Broadening out the definition to include more than idioms, it is common for researchers to make a stipulation that a word-string is formulaic when it cannot be generated using the regular form-meaning rules that create novel strings – that is, they locate formulaic material at the right-most end of the diagram in Figure 2. For example, in Erman & Warren's (2000: 32) definition, "one member of a prefab[5] cannot be replaced by a synonymous word without causing change of meaning or function and/or idiomaticity" (see also Wiktors-

5. For an exploration of the many different terms used for types of formulaic sequence, see Wray (2002a: 8ff).

son 2003; Forsberg 2006). Another formal property commonly associated with formulaicity under this view is grammatical oddity (e.g., *come a cropper; believe you me*).

Stipulating irregularity or non-transparency as the marker of formulaicity is a means of ensuring that all the examples identified definitely are formulaic. However, according to the morpheme-equivalence model, the definition is too conservative, because it excludes formulaic material that has not yet developed any oddities of form or meaning (see section 2.2).

2.4.4 *Are formulaic sequences always more than one word long?*

For some researchers it is nonsense to suggest that single words can be counted as formulaic. Yet a word, like a wordstring, can have a 'non-literal' meaning if it does not reflect the morphological components (e.g., *understand*), and/or if it has an irregular morphological composition (e.g., *children*). The morpheme-equivalence definition views formulaic sequences as behaving like single morphemes. It naturally follows that words doing likewise, and morphemes themselves, must be counted as formulaic.

A key practical advantage of accepting morphemes and words as formulaic is that it assists in functional analyses. It becomes possible to view *hallo*[6] as formulaic along with *nice to see you,* and *thanks* along with *thank you very much*. Similarly, it permits the inclusion of *into* along with *out of*, and of *well* along with *let me see*, which will be helpful with semantic analyses.

2.4.5 *Does code-switching respect the boundaries of formulaic sequences?*

One interesting development in relation to formal properties of formulaic language regards the case of code-switching. Backus (1999) predicts that, if formulaic sequences are holistically stored, then code-switching will never occur within formulaic sequences, only between them. This interesting proposal usefully demonstrates some of the challenges that can arise when taking a form-based approach to identification. Firstly, in order to test Backus' prediction it is imperative to have an independent means of identifying formulaic sequences – something that Wray & Namba (2003) attempt to provide (see later). Not having independently motivated criteria for identification would create a risk of circularity, since it will be very tempting to claim both that formulaicity determines where code-switching can take place, and that code-switching loci show us what is formulaic.

Secondly, important issues arise when one asks what it would mean if one **did** find wordstrings that appeared to be formulaic but also contained more than

6. Or indeed a cough or a raise of the eyebrows – also formulaic if they carry a reliable meaning attached to their form.

one language. Is it possible for a wordstring to be fundamentally constituted bilingually? Might someone not store whole a sentence like *it has a certain je ne sais quoi*? Finding such an example in a text would not need to imply that code-switching had taken place – it could have been learned that way. Frequent code-switchers, particularly if they interact with other code-switchers, might easily acquire mixed-code strings as lexical units.

Another reason why a formulaic sequence might feature more than one language would be if the formulaic element was a partly-lexicalized frame. Frames contain gaps for morphological detail or lexical insertions (e.g., *NP_i give [tense] NP_j a piece of PRO(NP_i)'s mind*, which can be realized as *I gave Mary a piece of my mind; the station master will give them a piece of his mind,* etc.). Accordingly, it becomes possible to envisage, in a code-switching situation, using one language for the formulaic frame, and the other language for the insertions (particularly lexical ones). Since the **frame** is viewed as formulaic, the change of language would not entail its termination, only its completion. Yet, of course, the outcome would be an alternation between the languages, contrary to the hypothesis that code-switching cannot occur until the end of a formulaic sequence. Navigating this issue entails addressing questions about formulaic language from within the framework of models of code-switching. For one attempt, see Namba (2008).

2.4.6 *Are formulaic sequences uncharacteristic of normal performance?*
Sometimes, particularly in relation to language learning (L1 and L2), it is possible to search for formulaic material on the basis of its being 'unusual for this person'. It might be precocious relative to normal output, if it was internalized holistically before the capacity to generate it from scratch developed. Alternatively, it might represent a throw-back to an earlier level of knowledge, if it incorporates an erroneous structure that the individual would no longer use in novel output. Such relative judgements require, of course, an accurate record of what else that individual can do (Myles 2004: 143).

2.4.7 *Can we identify formulaic sequences intuitively?*
Of all the potential approaches to the identification of formulaic sequences in text, intuition is probably the most troublesome. Most researchers recognize that intuition plays some sort of role in their approach, yet, of course, there must be a means of demonstrating that one person's intuitions are sufficiently like another's for conclusions based on them to be robust. Irrespective of whether intuition is judged a valid approach to identification, it is imperative that researchers indicate when they have used it. To bolster confidence in one's judgements, it can be useful to ask a range of native speakers to act as judges. Foster (2001), for example, required a minimal threshold of agreement between five out of seven native speaker judges, before an item was accepted as formulaic.

A: By my judgment there is something grammatically unusual about this wordstring.
B: By my judgment, part or all of the wordstring lacks semantic transparency.
C: By my judgment, this wordstring is associated with a specific situation and/or register.
D: By my judgment, the wordstring as a whole performs a function in communication or discourse other than, or in addition to, conveying the meaning of the words themselves.
E: By my judgment, this precise formulation is the one most commonly used by this speaker/writer when conveying this idea.
F: By my judgment, the speaker/writer has accompanied this wordstring with an action, use of punctuation, or phonological pattern that gives it special status as a unit, and/or he/she is repeating something just heard or read.
G: By my judgment, the speaker/writer, or someone else, has marked this wordstring grammatically or lexically in a way that gives it special status as a unit.
H: By my judgment, based on direct evidence or my intuition, there is a greater than chance-level probability that the speaker/writer will have encountered this precise formulation before in communication from other people.
I: By my judgment, although this wordstring is novel, it is a clear derivation, deliberate or otherwise, of something that can be demonstrated to be formulaic in its own right.
J: By my judgment, this wordstring is formulaic, but it has been unintentionally applied inappropriately.
K: By my judgment, this wordstring contains linguistic material that is too sophisticated, or not sophisticated enough, to match the speaker's general grammatical and lexical competence.

Figure 3. Wray & Namba's (2003) criteria for justifying intuitive judgements about formulaicity.

2.4.8 *Towards a solution for identification of formulaic sequences in text*

In the light of the discussion so far, it may seem unlikely that one could ever convincingly define, and reliably identify, all and only examples of formulaic language. Yet there is a practical need for methods that can allow research to progress. One way forward is to take decisions, but remain vigilant and reflective about what they assume and entail. To this end, Wray & Namba (2003) experiment with the potential for combining intuitive judgement with other approaches to identification. They invite the researcher first to examine the text and pull out instances that seem plausibly formulaic. Eleven criteria (Figure 3[7]) are then used as a means to establish the basis on which that judgement has been made. The

7. Full explanations of the criteria, and examples, are given in Wray & Namba (2003), Namba (2008) and Wray (2008).

criteria reflect a range of form-, meaning- and function-based approaches to definition and identification, but also allow for items that have no specific features marking them out as special, other than, perhaps, that they are known to have been said or heard before.

The Wray & Namba criteria cannot be used as a 'scoring system' for formulaicity because they are not cumulative: they are not all of the same order and some exclude others, so it does not follow that an item meeting three criteria is necessarily 'more formulaic' than one meeting only one criterion. Thus, rather than providing a fully robust basis for strong claims, the purpose of the approach is to assist researchers in (a) ensuring and demonstrating consistency between judges and across a dataset, (b) articulating the basis of their intuitions, and (c) gaining insight into possible biases in that intuition.

2.5 Embracing the opportunities

So far, the twin challenges of definition and identification have been problematized – and for good reason. The way to maximize credibility in formulaic language research is to employ the most robust definition capable of meeting the needs of the investigation. At the same time, it is important to understand the nature of the theoretical model underpinning any given definition. Deciding to include or exclude a particular type of example is sometimes just a means of keeping an analysis tidy, but in other instances it could jeopardize the logic of claims made about the phenomenon under investigation. In the discussion of frequency (2.4.1), for instance, it was noted how easy it is to omit informative examples by virtue of taking an over-conservative approach to identification.

By the same token, a theoretical model often has predictive power in several directions. We have seen how the morpheme-equivalence definition recognizes as formulaic items that are indistinguishable from novel constructions, and also single morphemes and words. But it makes other predictions too. If formulaic items can have different internal forms, and are formulaic on account of function and patterns in input, it follows that a broad range of communicative material could be deemed formulaic. In the remainder of this chapter some extreme examples of formulaicity will be used as a means of exploring aspects of the theory underpinning the morpheme-equivalence definition.

3. Boundaries

The morpheme-equivalence definition of formulaicity proposes that formulaic sequences are learned whole and stored whole, with a reliable meaning attached to the form. Holistic form-meaning mapping explains how pragmatic meaning

can be associated with a complete string, independently of its components. For instance, *don't do anything I wouldn't do!* carries, in some circumstances, connotations of doing precisely the opposite, and in others is little more than a friendly valediction. However, the theory enables other, stronger claims as well. Wray (e.g., 2002a: 94f; Wray & Perkins 2000) suggests that speakers use the holistic message feature of formulaic sequences to constrain the hearer's thoughts and reactions, and thereby to direct the hearer into a particular, desired perception or response. Would those constraints extend also to the speaker? That is, can formulaicity affect the capacity of a speaker to express him or herself freely?

In order to answer the question 'Do formulaic sequences constrain expression?' three situations will be considered. In the first, the user is not **required** to be formulaic, but using non-formulaic material exacts a price. At what point, and in what circumstances, will a language user find the expressive constraints of formulaic sequences sufficiently uncomfortable for the more costly alternative to be preferable? The second situation is one in which formulaic sequences are deliberately imposed as means of preventing the speaker and hearer realising that anything else could be said: to what extent can speakers and hearers be so-controlled? Finally, in the most extreme situation of all, there is no alternative to formulaicity. What happens when one has more to say than the communication system permits? By asking these questions, it will be possible to explore the nature of formulaicity in new ways.

3.1 Escaping formulaicity, but at a price

Two types of investigation assist in investigating how formulaic language constrains expression when it doesn't **have** to. The first entails an augmentative communication (AC) software program, TALK (e.g., Todman, Rankin & File 1999). AC systems convert typed input into computer-generated speech, and a particularly feature of TALK is its ability to increase the production speed of such output to a level that makes conversational exchange possible. It is possible because the user, rather than inputting text in real time, pre-stores what she expects to need, and selects the material simply by clicking on an icon, making the link from choice about what to say to production much faster.

Anticipating what one will need to say in one's own half of a conversation, when one does not know what the other person will say in response, might appear a very unsatisfactory way to manage conversation. However, an experienced user can employ TALK very effectively (Wray 2002b). Of interest to us here is the extent to which one user, Sylvia, was content to operate within the constraints of prefabricated material, and how she navigated her communicative activity so as to minimize the limitations of doing so.

The second source of evidence is a research project inspired by TALK, in which the pre-storage method was applied to language learning. Non-native speakers of English were asked to think ahead to conversations they knew they were going to have, and guess what they would need to say. Their ideas were re-expressed by a native speaker, and the nativelike versions were audio-recorded for them to memorize and practise. For full details of this research, see Fitzpatrick & Wray (2006) and Wray & Fitzpatrick (2008).

In both TALK and the language learning experiment, the individual had much to gain by relying on the formulaic material: it enhanced fluency and accuracy respectively. But in both cases there was an escape from formulaicity if necessary. The language learners had a general knowledge of English, so they could abandon what they had memorized and construct something new. They could also edit a previously memorized string, to make it fit their present needs better. The disadvantage was that they abandoned what they knew to be nativelike, and reverted to their own, non-nativelike constructions. Meanwhile, Sylvia also had two options if she did not want to use what she had pre-stored in TALK. She could create new utterances in real time, or edit a pre-stored item. However, both were time-consuming and communicatively disruptive procedures. The question of interest, then, is whether, and at what stage, the desire to express a specific idea that had not been prepared would override the desire to remain fluent and/or nativelike.

The language learners **very** easily departed from the prepared material, both unintentionally and deliberately. The unintentional changes suggest that it was difficult for them to trust holistically-stored material. The deliberate changes sometimes altered facts – such as the time that a meeting would take place – so that they were trading formal accuracy for factual accuracy. However, the learners also made changes when they felt that the nativelike expression did not convey their non-native thoughts adequately. They would rather be true to their thoughts and perceptions than sound nativelike (Fitzpatrick & Wray 2006).

In contrast, although Sylvia would spell out a novel response if she absolutely had to, she prioritized sticking with pre-stored material whenever she could, to retain the flow of expression. She minimized the disadvantages of this choice by employing strategies for coping with poor matches between what she wanted to say and what she could say, including using fillers like *I haven't thought about that much* or *That's a good question*. She was also prepared for her statements to be untrue – she valued fluency over factual accuracy – and she often gave interpretative responsibility to the hearer, who needed to apply pragmatics to make what she said match the context, e.g., using the pre-stored *I like shopping in Dundee* as a response to *Where did you go shopping?* rather than creating *In Dundee*.

In these two investigations, it seems that expression was not compromised unduly by the advantages of formulaicity, though Sylvia, perhaps on account of lengthy practice, was able to hold out longer than the learners before she switched to novel expression.

3.2 External attempts to control expression and thought

The second scenario entails attempts to use formulaic language for political and social control. George Orwell (1946; 1949) had strong views about the capacity of formulaic language to compromise incisive thought:

> ready-made phrases … will construct your sentences for you – even think your thoughts for you … and at need they will perform the important service of partially concealing your meaning even from yourself (Orwell 1946: 135).

Orwell was suspicious about how dictatorships might use and mould language to control a population. Under the totalitarian regime portrayed in his novel *Nineteen Eighty-Four* (1949), saying the wrong thing was highly dangerous. In order to avoid being betrayed to the thought police, people learned to speak only the prescribed slogans, until the ideas conveyed in the slogans became their only thoughts.

Aspects of Orwell's fictional world were, according to Ji (2004; see also Ji, Kuiper & Shu 1990), translated into reality as 'linguistic engineering' during the Cultural Revolution in Maoist China. She describes this as "the great attempt to produce new, revolutionary human beings by enforcing the constant repetition of revolutionary formulae" (2004: 317). The aim was to "enforce the habitual use, in relevant contexts, of numerous fixed expressions and standardized scripts that embodied 'correct' attitudes or that had 'correct' propositional content" (p. 4). Slogans and quotations from Mao's writings were part of the fabric of everyday interaction, so that "their message would sink into people's brains and guide their behaviour" (p. 5). Mao's rationale was that "If people could be made to speak formulaically and through that learn to think formulaically … all individuality, all merely personal aspirations would be destroyed" (p. 178).

How could people be persuaded to adopt this kind of formulaic expression, and formulaic thought? According to Ji it was achieved by confusing people about how they were positioned relative to the regime, so that they felt insecure about their own behaviour. By repeatedly altering the definition of what constituted pro- and anti-regime attitudes and actions, Mao made it difficult for individuals to remain sure whether they were behaving appropriately (Ji 2004: 143). As a result, people felt vulnerable to making unintentional slips in what they said, so that "the safest course … was to speak and write in Chairman Mao's own words"

(p. 155). Mao's *Little Red Book* thus became an object of study and memorization, and social and political identity was directly signalled through the use of certain approved sayings, reproduced formulaically.

Is it actually possible to control a huge population through the imposition of formulaic language? Orwell clearly thought it was. However, for Ji (2004) the Chinese case suggests otherwise. The reason lies in the way a communal belief needs to be translated into individual action. Restating prescribed tenets is one thing, but acting them out in one's own life requires the tenets to be interpreted and applied to new situations. It is insufficient only to have a limited set of responses. To interact effectively with novel situations one must know how to interpret and extend the formulaic material, and one must use pragmatics to extrapolate the meaning of the old to the new. In this way, formulaicity comes into contact with creativity, and the necessity of this juxtaposition entails that individuals maintain the capacity for novel thought. Novel thought, and novel language, therefore are needed for the extension and consolidation of formulaic behaviour, while, of course, also being the means by which subversion arises (for a more detailed discussion, see Wray 2008).

3.3 Absence of novelty

Suppose there is no escape from formulaicity at all: what then happens to novel expression? This scenario has been explored as a stage in the evolution of language (e.g., Wray 1998, 2000, 2002c; Mithen 2005), but we shall focus here on more recent history, when there would have been a clash between the flexibility of human language and the severe constraints of the signalling system used in its place.

Trumpet and bugle calls[8,9] were introduced from the 13th Century, and were used operationally in battle until WW1, when radios were introduced. In UK camps and barracks, the signals remained the most effective way to instruct dispersed individuals until, in the 1970s, soldiers were issued with wristwatches and daily schedules. An account by General Custer's wife Elizabeth of life in a cavalry

---

8. Information in this section comes primarily from Powell (2000a,b) and personal interviews with trumpeters, including David Edwards and Crispian Steele-Perkins. A more detailed account of the research into military signalling is in Wray (2008).

9. The first signals were on drums, but they became impossible to hear once cannons were introduced in warfare. The long trumpet first took over, but it was heavy and unwieldy. The invention of the wound trumpet and, ultimately, the bugle, made the instruments more portable. Trumpets continued in use in cavalry regiments, since the rider did not have to bear the weight other than when playing.

regiment in the third quarter of the 19th Century (Custer 1890) gives something of the flavour of how signals operated:

> [The trumpet] was the hourly monitor of the cavalry corps. It told us when to eat, to sleep, to march, and to go to church. Its clear tones reminded us, should there be physical ailments, that we must go to the doctor … We needed timepieces only when absent from garrison or camp. The never tardy sound calling to duty was better than any clock and … we found ourselves saying 'Can it be possible? There's 'Stables', and where has the day gone? (p.v).

Signals were holistic, without meaningful component parts,[10] and often had very complicated meanings, e.g., "[B]y a 'Troop' [you must] understand to shoulder your muskets, to advance your pikes, to close your ranks and files to their order, and to troop along with or follow your officer to the place of rendezvous or elsewhere" (Custer 1890: x). In battle, buglers and trumpeters not only sent orders from the commander but also undertook reconnaissance, so that a great many different messages might potentially be needed. Yet, of course, there were constraints on how many signals could exist, since they needed to be learned by the players and interpreted by a wide range of hearers. Signals used too rarely would be forgotten.[11] The holistic nature of the signals meant, also, that there was no scope to modify signals to convey variable information. How could the gap between expressive need and practical provision be bridged? Would the finite set of signals constrain the messages that could be sent? Or would the need to send novel messages be sufficient to subvert the signalling system?

The pressures on the system were evidently very real, for a number of changes were made over time. Firstly, the number of signals increased, as a means of encoding many very precise messages. By the early 19th Century there were specific calls meaning such things as *the enemy has infantry and cavalry* and *the enemy's cavalry is advancing*. Secondly, the same signal was sometimes used for two related meanings, leaving the hearer to apply pragmatics to determine which was intended. For example, one signal meant both *draw your swords* and *return your swords*. Thirdly, a very particular kind of 'grammar' was introduced in the British military in 1835, with regimental and brigade calls. These signature tunes preceded another signal, to indicate which unit of men the order was

---

10. There were minor exceptions to the arbitrariness, e.g., the instruction to gallop was a faster tune than the instruction to walk, though the notes of the tune were also different.

11. Compare Wray (2002c: 120), discussing the same constraints on holistic signals in human protolanguage.

addressed to. Finally, calls could be concatenated into a macro-call that gave a sequence of instructions, as in Figure 4. Although such macro-calls superficially resemble compositional language, they contained no grammar (other than the signature call), and there was a direct relationship between temporality and linear order. Therefore, a more appropriate analogy would be sentences contributing to a narrative.

What is striking is that more sophisticated grammatical relations were not introduced. In the absence of a means of encoding, for instance, that **although** call A was true, **nevertheless** so was call B, a great deal needed entrusting to the pragmatic interpretative powers of the hearer.

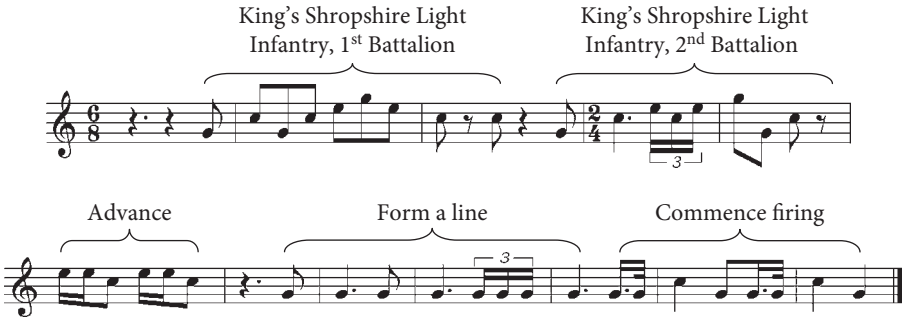


Figure 4. A composite trumpet call (after 1835).

Under these constraints it can be seen that the signalling system itself could not supply an entirely effective bridge between the limits of prefabrication and the need to be novel. Yet, by the 19th Century, the Napoleonic, Crimean and American Civil wars employed sophisticated weaponry and complex battle plans. Did commanders mould their strategy to accommodate the limitations of signalling? Inevitably, situations must have arisen in which, with no alternative available, the signals determined the command. However, there were ways of avoiding it. Firstly, the commander could lay out a plan for a certain hypothetical scenario, and instruct his officers in advance about what to do. Should the situation arise for it to be enacted, he could send the signal meaning *carry out the order*. Of course, this strategy only worked for something that had been foreseen and agreed. For other situations, messengers were sent with verbal instructions. In short, even where the signalling system did not have any means of encoding novel messages, so that it was maximally constrained, the human capacity for novel thought and novel expression was rarely defeated.

3.4 Evidence from the boundaries

In all of these examples the formulaicity of the limited system was in conflict with the need for creativity in expression. In each case, the user struggled to manage communicating without easy recourse to novel material, and exercised choice about when to abandon formulaicity in order to release new meaning. The costs of abandoning formulaicity varied from situation to situation – Sylvia had to sacrifice the advantages of fluent conversation whenever she produced novel material; the language learners relinquished the opportunity to sound nativelike; those under Mao's regime risked making a mistake that could have serious consequences for their lives; and commanders who dispatched a messenger risked losing both him and the message.

The discovery that, in every case, there was a point when formulaicity would be abandoned, illuminates our understanding of what happens at the more everyday boundary between formulaic and novel language. The evidence suggests that humans will strive always to retain the creative edge. There is an inherent tension between the opportunities offered by formulaic language – savings in processing, fluency, sounding like others, shortcuts to the most useful linguistic material, the manipulation of others – and the necessity to retain creativity and novelty in order to hone the delivery of messages to each specific circumstance. The balance between the two must be viewed as fluid, and as contingent on context, culture and need (Wray & Grace 2007).

4. Conclusion

It may seem an odd juxtaposition, first to expound the issues surrounding appropriate approaches to the definition and identification of formulaicity, and then to ask questions about formulaicity using examples that some would consider well outside the range of useful definition. However, this dissonance was very much the point of the exercise. **Should** sentences pre-stored in a computer or in the human memory be considered instances of formulaic language or not? Should the adoption of prescribed forms of words for socio-political reasons be viewed as part of the scope of formulaic language use? And can signals that are not even linguistic, but musical, be seen as part of the same phenomenon as formulaic language? The answer depends on the theoretical basis upon which one conceptualizes formulaicity. The morpheme-equivalence model proposes that benefits can ensue from using a code with an internally complex form as if it were not internally complex. With most normal language – true idioms are the exception – it is difficult to

test the claim that something really is being handled formulaically, because of the hair-trigger option of engaging with the parts rather than the whole. The extreme examples used in this chapter test the case by altering the balance of desirability and/or capacity to switch to a more analytic approach to encoding and decoding. These extreme examples seem to operate on the same axes as more ordinary formulaic language. Nevertheless they are hardly prototypical. It is by exploring the ways in which the differences from the prototype distance them in terms of their treatment in communication that we can home in on the detail necessary for establishing effective ways of identifying formulaic sequences in the material we choose to analyse.

References

Backus, Ad. 1999. Evidence for lexical chunks in insertional codeswitching. In *Language encounters across time and space,* B. Brendemoen, E. Lanza & Else Ryen (Eds), 93–109. Oslo: Novus Press.

Conklin, Kathy & Norbert Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29: 72–89.

Custer, Elizabeth B. 1890. *Following the guidon*. New York NY: Harper & Brothers. (Republished 1998 by Digital Scanning Inc, Scituate, MA, www.digitalscanning.com)

Ellis, Nick, Rita Simpson-Vlach & Carson Maynard. 2006. The processing of formulas in native and second-language speakers: Psycholinguistic and corpus determinants. Paper presented at the International Conference on Exploring the Lexis-Grammar Interface, Hanover, Germany, 5–7 October.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20: 29–62.

Fitzpatrick, Tess & Alison Wray. 2006. Breaking up is not so hard to do: Individual differences in L2 memorization. *Canadian Modern Language Review* 63(1): 35–57.

Forsberg, Fanny. 2006. Le langage préfabriqué en français parlé L2. Ph.D. dissertation, Stockholm University.

Foster, Pauline. 2001. Rules & routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching pedagogic tasks: second language learning, teaching and testing*, M. Bygate, P. Skehan & M. Swain (Eds), 75–97. London: Longman.

Ji, Fengyuan. 2004. *Linguistic engineering: Language and politics in Mao's China*. Honolulu HI: University of Hawai'i Press.

Ji, Fengyuan, Koenraad Kuiper & Shaogu Shu. 1990. Language and revolution: Formulae of the Cultural Revolution. *Language in Society* 19: 61–79.

Jones, Martha & Sandra Haywood. 2004. Facilitating the acquisition of formulaic sequences. In *Formulaic sequences: Acquisition, processing and use*, Norbert Schmitt (Ed.), 269–300. Amsterdam: John Benjamins.

Mithen, Steven. 2005. *The singing neanderthals*. London: Weidenfeld and Nicolson.
Myles, Florence. 2004. From data to theory: The over-representation of linguistic knowledge in SLA. *Transactions of the Philological Society* 102(2): 139–168.
Namba, Kazuhiko. 2008. English-Japanese bilingual children's code-switching: a structural approach with emphasis on formulaic language. Ph.D. dissertation, Cardiff University.
Orwell, George. 1946. Politics and the English language. *Horizon: A Review of Literature and Art,* 64(April). (Reprinted in *The collected essays, journalism and letters of George Orwell,* Vol IV (1945–1950), S. Orwell & I. Angus (eds.) 1968, 127–140. London: Secker & Warburg)
Orwell, George. 1949. *Nineteen eighty-four*. London: Secker & Warburg/Penguin.
Peters, Ann M. 1983. *Units of language acquisition.* Cambridge: CUP.
Powell, Richard. 2000a. The bugle's tale (part 1). *Band International: Journal of the International Military Music Society* 22(1): 16–17
Powell, Richard. 2000b. The bugle's tale (part 2). *Band International: Journal of the International Military Music Society* 22(3): 114–118
Todman, John, David Rankin & Portia File. 1999. The use of stored text in computer-aided conversation: A single-case experiment. *Journal of Language and Social Psychology* 18(3): 320–342.
Underwood, Geoffrey, Norbert Schmitt & Adam Galpin. 2004. An eye-movement study into the processing of formulaic sequences. In *Formulaic sequences. Acquisition, processing and use*. Norbert Schmitt (Ed.), 153–172. Amsterdam: John Benjamins.
Van Lancker, Diana, Gerald J. Canter & Dale Terbeek. 1981. Disambiguation of ditropic sentences: acoustic and phonetic cues. *Journal of Speech and Hearing Research* 24: 330–335.
Wiktorsson, Maria. 2003. *Learning idiomaticity: A corpus-based study of idiomatic expressions in learners' written production* (Vol. 105). Lund: Department of English, Lund University.
Wray, Alison. 1998. Protolanguage as a holistic system for social interaction. *Language & Communication* 18: 47–67.
Wray, Alison. 2000. Holistic utterances in protolanguage: The link from primates to humans. In *The evolutionary emergence of language: Social function and the origins of linguistic form,* Chris Knight, Michael Studdert-Kennedy & James R. Hurford (Eds), 285–302. Cambridge: CUP.
Wray, Alison. 2002a. *Formulaic language and the lexicon*. Cambridge: CUP.
Wray, Alison. 2002b. Formulaic language in computer-supported communication: Theory meets reality. *Language Awareness* 11(2): 114–131.
Wray, Alison. 2002c. Dual processing in protolanguage: competence without performance. *The transition to language,* Alison Wray (Ed.), 113–137. Oxford: OUP.
Wray, Alison. 2004. 'Here's one I prepared earlier': formulaic language learning on television. In *Formulaic sequences. Acquisition, processing and use*, Norbert Schmitt (Ed.), 249–268. Amsterdam: John Benjamins
Wray, Alison. 2008. Formulaic language: Pushing the boundaries. Oxford: OUP.
Wray, Alison & Tess Fitzpatrick. 2008. Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In *Phraseology in foreign language learning and teaching*, F. Meunier & Sylviane Granger (Eds), 123–147. Amsterdam: John Benjamins.
Wray, Alison & George W. Grace. 2007. The consequences of talking to strangers: Sociocultural influences on the lexical unit. *Lingua* 117(3): 543–578.

Wray, Alison & Kazuhiko Namba. 2003. Formulaic language in a Japanese-English bilingual child: A practical approach to data analysis. *Japan Journal for Multilingualism and Multiculturalism,* 9(1): 24–51.

Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20(1): 1–28.

PART II

Structure and distribution

Formulaic tendencies of demonstrative clefts in spoken English

Andreea S. Calude
The University of Auckland*

Abstract

Despite having been noted as a frequent construction in spoken language, little has been said about the demonstrative cleft. Clefts such as *that's what I am talking about*, or *that's what I mean*, are not entirely fixed in their structure; however, they do exhibit recurring patterns and "preferred formulations" (Wray 2006: 591). An investigation of demonstrative clefts in excerpts of spontaneous conversations from the Wellington Corpus of Spoken New Zealand English shows that, aside from being the most frequent cleft type in conversational English, the cleft is characterized by structural fixedness, fluency, and non-salient reference. As claimed by Ford, Fox and Thompson, grammar is (in general) "a collection of crystalizations of routines" (2002: 120); and nowhere is the emergent (Hopper 1987, 2001) nature of grammar more clear than in spoken language. The demonstrative cleft is an example of such a routine, and thus worth investigating further.

1. The demonstrative cleft construction

This paper reports findings related to the construction of demonstrative clefts (termed after Biber et al. 1999: 962), as exemplified below in bold in (1)[1] and (2).

---

* The University of Reading

1. The Wellington Corpus of Spoken New Zealand English contains tags for various discourse features, such as pauses, laughter, latches and so on. Some of these have been left out

(1) AQ: should be hokey pokey mokey
BG: okay so what RATE i mean do you want twenty dollars an hour or
→ AQ: well just about the normal five or six five or whatever you **that's what you usually pay for everything isn't it** round the house ⟨,⟩
BG: yeah five dollars an hour?
→ AQ: **that's what you said** you said ⟨unclear word⟩
BG: yeah but then you could take hours couldn't you you could take hours to do something
AQ: oh well as if I would (WSC, DPC089:0245-0285)[2]

(2) BA: oh ⟨laughs⟩
XX: is that a dynamo thing
TR: dymo yes ⟨latch⟩
XX: dymo
TR: I found my dymo my tape my not negotiable ⟨12: 00⟩ stamp for my cheques
XX: very handy
→ TR: **THAT'S what I thought**
XX: what are you going to use your dymo for
TR: I don't know ⟨latch⟩
XX: you can put your name on your door now that I've got my name on mine
(WSC, DPC025:1610-1665)

An inspection of excerpts of unplanned spontaneous conversation (circa 200,000 words) from the Wellington Corpus of Spoken New Zealand English[3] (henceforth WSC) shows that demonstrative clefts exhibit formulaic tendencies. They have a relatively fixed structure, allowing only a narrow range of elements to occur in their slots, a distinct function in discourse as organizational and discourse-managing markers, and they are associated with a specific type of data (i.e., informal spoken language).

It is perhaps of no surprise to see that a construction which has been associated with informal speech (Biber et al. 1999) exhibits formulaic tendencies. Work by Aijmer 1996; Biber et al. 1999; Biber & Conrad 1999; Miller 1994; Miller & Weinert 1998; Thompson 2002; and others shows that spontaneous

---

of the examples given here for ease of comprehension. The remaining ones included and their meanings are given in Appendix A.

2. The examples cited from the WSC contain information pertaining to the file used (i.e., DPC089 stands for file 89 of conversation data), and the time in the transcript where the language excerpt comes from (i.e., 0245–0285 indicates a portion of discourse from 245 seconds into the recording until 285 seconds).

3. See Holmes, Vine & Johnson (1998) for a guide to the corpus and the corpus website at the Victoria University of Wellington (http://www.vuw.ac.nz/lals/corpora/index.aspx#wsc).

spoken language involves heavy use of "prefabricated chunks", "lexical bundles", "conversation routines" and "formulaic expressions".

The paper is organised as follows. First, the background of the construction of demonstrative clefts is discussed in Section 2. We will see that despite being noted for its frequent use in informal, spoken language, little is known about the construction. Adding to that, researchers have previously classified it together with reversed *wh*-clefts, which has created difficulties and confounded results for both cleft types. The problems come about from the fact that in some studies (specifically those analyzing spoken language), researchers reporting on reversed *wh*-clefts are in reality reporting on demonstrative clefts since most of their data is made up of the latter (which they have grouped together with the former), for example Miller & Weinert 1998; Herriman 2004; and Oberlander & Delin 1996. Section 3 presents the formulaic tendencies of demonstrative clefts found in the WSC excerpts. These have to do with structural, phonological and discourse-related properties of the cleft construction. The paper concludes with a brief summary section.

2. Background

Cleft constructions are sometimes described as being the "result" of a simple clause[4] which is "cleaved" and re-arranged inside a particular schema (depending on the cleft type), for the purpose of highlighting a given constituent (typically, its subject or its object). Typically, discussions of clefts in English involve the schemas of *it*-clefts or of *wh*-clefts (basic or reversed), as given in (3), compiled from Huddleston and Pullum (2002: 1414–1427).

(3) a. *It*-cleft: It + BE + foregrounded element + REL CL
b. *Wh*-cleft: (basic) WH-word + foregrounded element + BE + REL CL
(reversed) Foregrounded element + BE + WH-word + REL CL

For instance, the sentence given in (4) can be turned into a cleft (*it*-cleft or *wh*-cleft), highlighting the subject *the alligator*, or the prepositional object *in a shallow pond*, as given in (5a–f).

(4) The alligator lives in a shallow pond.

4. As mentioned by Huddleston and Pullum (2002: 1422), complex sentences can also be turned into clefts, e.g., *It was staying at home and being left behind that worried her most*, but these are rare in general, particularly in spoken data, and not attested in the excerpts of conversation from the WSC corpus.

(5) a. It is **the alligator** that lives in a shallow pond. (*it*-cleft, focusing subject)
b. **The alligator** is what lives in a shallow pond. (*wh*-cleft, focusing subject)
c. That which lives in a shallow pond is **the alligator**. (reversed *wh*-cleft, focusing subject)
d. It is **in a shallow pond** that the alligator lives. (*it*-cleft, focusing object)
e. **In a shallow pond** is where the alligator lives. (*wh*-cleft, focusing object)
f. Where the alligator lives is **in a shallow pond**. (reversed *wh*-cleft, focusing object)

Cleft sentences contain three major components: the **cleft constituent**, which is the phrase or clause that is being highlighted or focused, the **copula** verb (in English, *be*), and the **cleft clause**, containing the remaining parts of the uncleaved sentence (*it*-clefts also involve a fourth element, namely the cleft pronoun *it*).

A different way of looking at clefts is to think of them as pairings of a variable (expressed by the cleft clause) and its associated value in a particular context (given by the cleft constituent). So for instance, in (5a-c), the variable is the entity which lives in a shallow pond, and its value is the alligator. Similarly, in (5d-f), the variable is the place where the alligator lives, and its associated value is the shallow pond.

A third schema, alongside those of *it*-clefts and *wh*-clefts has been pointed out in Calude (2008), namely that of demonstrative clefts, given below, and exemplified in (6a-c).

(6) Demonstrative-cleft: Demonstrative pronoun + BE + [*wh*-word + REL CL]
a. That's what I don't like about her.
b. That's what he thought.
c. This is what it's all about.

The formula shown in (6) is strikingly similar to that of reversed *wh*-clefts, seen earlier in (3b). This similarity explains the overwhelming number of studies which have placed demonstrative clefts under the umbrella of reversed *wh*-clefts (Collins 2004, 1991; Hedberg 2000; Herriman 2004; Lambrecht 2001; Miller & Weinert 1998; Oberlander & Delin 1996, and Weinert & Miller 1996). However, this paper is not concerned with clarifying the reasons why investigating demonstrative clefts in isolation from reversed *wh*-clefts proves to be worthwhile – this is discussed in detail in Calude (2008). Here, the discussion will be limited to only briefly touching upon the main difference between the two clefts, namely the deictic properties of the demonstrative cleft, which are not present in reversed *wh*-clefts.

First, it is perhaps worth noting that what is at issue is not so much whether the demonstrative cleft is sufficiently distinct from reversed *wh*-clefts to warrant a separate label (or a class of its own). Instead, the question is whether there is anything to be gained from the separate investigation of the construction; that is, whether there is something new and interesting to be learned which may otherwise be

overlooked. The answer to the question, as proposed in Calude (2008) and echoed here, is that yes, there is indeed something to be gained from such an analysis.

As mentioned earlier, the difference between the two cleft constructions has to do with the fact that unlike reversed *wh*-clefts, demonstrative clefts have strong deictic links to the surrounding discourse, arising from the use of the demonstrative pronoun in cleft constituent position. These deictic links have implications for the syntactic as well as information structure of the demonstrative cleft. Although the arguments for analysing demonstrative clefts separately from reversed *wh*-clefts are presented in Calude (2008) and thus beyond the scope of the current paper, one example is given below, in order to give a taste for how the two clefts differ.

Compare the demonstrative cleft given in example (7a) with the reversed *wh*-cleft in (7b).

(7) a. OR: so one of those has got to come and make ME up eventually but what's hap but Gareth's opinion is the same as mine in reality is that you go out and you take somebody out in the back yard ⟨5: 00⟩ maybe for a start off and spend a couple of hours or so going over all the basics of things so until what we'd like to call it a preliminary certificate's passed where we know that YOU can operate the ladder and from then on it should be able to I should just be able to just go out with you and nobody else on the ladder

WL: until yeah mm

OR: and do a training

WL: yep

OR: I want because the only way you get to know how to work the thing is by getting out there and working it and the ⟨latch⟩

WL: yeah ⟨latch⟩ worst thing about it is with the senior station officer being the instructor is that the pump crew's got to go everywhere the ladder is while we're doing our training ⟨sighs⟩ yeah oh yeah like I was going to say because that changed with mind you when snoopy was made up to instructor cos thumper just used to go out and um ⟨latch⟩

OR: with snoopy ⟨latch⟩ with snoop and if they had a fire call he just drove off
→ to the fire call yeah well **that's what we hopefully we will be able to do** although it's going to be a hassle with me what I want to do is get clearance from Clark to say (WSC, DPC291:0100-0170)

b. AC: Ann's not the social work she'd be a disaster

BS: right

AC: Ann's cultural affairs which means you organise a garden party which
→ means NOTHING basically **cultural affairs is what they put the person who's not going to do any work on** ⟨laughs⟩so they put Ann on there she sort of suits it though cos Ann doesn't like everyone thinks that Ann organised the ski trip but her PARENTS organized the WHOLE THING it's true (WSC, DPC059:0640-0670)

In the reversed *wh*-cleft uttered by speaker AC, *cultural affairs is what they put the person who's not going to do any work on*, the value which the cleft clause references (namely, the noun phrase *cultural affairs*) is contained within the cleft construction. The hearer does not need any further information in order to make sense of the cleft. This value is a simple noun phrase, and clearly, it is uttered by the same speaker as the one producing the reversed *wh*-cleft, and occurs in the same turn as the cleft (since it is actually part of it).

In contrast, in the demonstrative cleft uttered by speaker OR, *that's what we hopefully we will be able to do*, the cleft constituent *that* refers to the entire portion of discourse mentioned earlier by OR regarding going over all the basics, and introducing the training schedule for what would be termed the preliminary certificate. The value of the cleft is still expressed by the same speaker as the one producing the cleft, but it occurs several turns prior to it, and consists of several clause complexes.

It turns out that speakers make use of demonstrative clefts not only to point to their own previous (or upcoming) speech, but also to the speech of other participants present at the time of interaction. Cleft constituents are used to reference single phrases, full clauses, or entire portions of discourse spanning several complex clauses (as exemplified in 7). Finally, the material being referenced by the demonstrative pronoun can occur in close proximity to, but also in isolation from, the cleft construction (i.e., several turns prior to it).

Given these possibilities, the majority of demonstrative clefts, however, are primarily used to refer to the same speaker's previous contribution (anaphoric), which consists of an entire clause or a longer, more complex portion of discourse (extended value[5]), uttered in the same turn as the cleft construction.

Additionally, demonstrative clefts play a distinctive role in organizing and managing the flow of conversation. Reversed *wh*-clefts are said to have a summative, "remind-me" role in discourse (Collins 2004: 70; Miller 1996: 113), pulling the discourse together at the end of a topic strand, or adding "newsworthy comments which highlight modal meaning of volition or necessity or explanations of causal relations", cf. Herriman (2004: 466).

Demonstrative clefts are used for a different purpose. While they have varying functions, one common thread shared by them is that of regulating the discourse whether by signaling how previously mentioned ideas relate to each other, or by highlighting salient entities in the discourse, or further still by enabling speakers to take the floor in a non-threatening fashion, or finally, by encouraging the current speaker to continue talking (see Calude 2008, for details and examples).

5. Following Collins (1991), the values being referenced by the demonstrative clefts are extended, if expressed by a single phrase, and non-extended, if coded by one or more clauses.

A secondary role of the demonstrative cleft is that of providing explanations and clarifications of previous sections of the discourse. This is not a discourse management role in the same way as the other functions. However, it does contribute to the overall comprehension of the discourse and it allows speakers to make sense of each other's contributions.

The deictic links which demonstrative clefts have with surrounding discourse, and the organizational role they play in managing it make the cleft construction an economic and efficient tool for speakers. Economy comes from the cleft's low informative content, and its allowing speakers to elaborate on previous material by making reference to it (via demonstrative pronouns), and thus bypassing the need to integrate it in complex, cognitively-demanding structures. Efficiency comes as a result of being able to refer to previous discourse, without the need to repeat or restate it. Finally, unlike (planned) written language, (spontaneous) spoken language is unstructured and loose, predisposed to sudden changes in topic, digressions and topic re-instatements. There is no paragraphing or set way of organizing the flow of ideas. Hence, there is a greater need for discourse markers and specialized constructions such as demonstrative clefts, which help orient participants in discourse, and help them make sense of what is coming up and how it relates to what has already been discussed.

Having discussed some of the major characteristics of demonstrative clefts, the implications of their deictic properties and also their distinctive function in discourse, we now turn our attention to their formulaic tendencies.

3. Evidence of formulaic tendencies in demonstrative clefts

Formulaic aspects of human languages have always played a role in linguistics theory, with researchers recognizing the presence of idioms – one of the (if not the) most formulaic construction there is. However, against the dominant backdrop of work focusing on the creative faculty of language, linguists have neglected the strong presence of formulaicity in language use. Depending on how we define the notion of a FORMULAIC EXPRESSION, we may find that our speech involves as much as 80 per cent formulaicity (Altenberg 1998: 102).

This point brings us to several issues which still remain unclear and require further attention, namely the problem of defining and that of detecting formulaic language. A widely used definition of a formulaic sequence comes from Wray's seminal 2002 book:

> "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002: 9).

While it is recognized that there are problems with this definition, in that it is almost impossible to know with absolute certainty whether or not a sequence is actually stored and retrieved whole from memory, there are also some merits in its inclusiveness, allowing various degrees of formulaicity to be accounted for (such as sequences which are only partly fixed and which incorporate some novel choices in their structure). It is perhaps fair to say that the definition does not help us in detecting formulaicity, but merely aids in painting a picture of the kind of phenomenon we are looking at.

The history of the study of formulaic language phenomena has been such that various researchers working in different areas (be it in first or second language acquisition, psycholinguistics, syntax, corpus linguistics, etc) and using different criteria and terminology were essentially looking at the same thing (admittedly, with some subtle differences). One consequence of this divide is that we have ended up with a wealth of different terminology (see Wray 2002: 9 for a sample). Another is that, thankfully, we have also been fortunate enough to inherit a plethora of diverse criteria for identifying formulaic expressions. This is indeed fortunate, since ideally, we would like to see more than just one or two pieces of evidence suggesting formulaicity in order to be confident that we are indeed dealing with a formulaic construction (Read and Nation 2004). Multiple pieces of evidence are needed because, as we will see below, some of the properties used for identifying formulaicity (such as, for instance frequency counts) are not exclusively found in formulaic expressions.

Wray gives a summary of some of the ways in which we may detect formulaicity in her 2002 book, Chapter 2. These include structural, semantic and phonological criteria, frequency counts and evidence from second language use.

Frequency counts are an obvious choice for identifying formulaicity, despite the fact that frequent expressions need not be formulaic (*the man*), and conversely, many which are formulaic are not frequent (*The King is dead, Long Live the King, kick the bucket*). Furthermore, the same expression may be formulaic in one context, but not another, e.g., *keep your hair on* in the sense of wearing a wig is not formulaic, but not getting stressed or upset is (Wray 2002: 25–31). As far as structure is concerned, formulaic expressions are fixed (to various degrees, in various ways), disallowing other possible elements to occur inside the formula (?*he put his left foot in his mouth*) and also potentially fossilized grammatically (*if I were/was you*) or immune to the rules of grammar (*by and large*). Furthermore, their meaning can be opaque, as in *selling oneself short* (which does not really have anything to do with selling). Since formulaic expressions are treated as a unit, they share some properties with single words. For instance, there are often no pauses or breaks during the course of a formulaic expression. Similarly, if participants engage in code-switching, this is likely to happen at the

boundary of formulaic expressions, not during the course of these (Backus 1999). Finally, some formulaic expressions are associated with particular data types (e.g., the speech of auctioneers, sports commentators, formal letters of rejection etc). We will see more details about some of these criteria below, where they are discussed in relation to demonstrative clefts.

What is crucial about the above criteria, however, is that not all properties apply to all formulaic expressions. Typically, they will only exhibit a sample of these characteristics, and obviously, a higher number of properties observed correlates with a higher degree of formulaicity in the expression. This view of formulaicity understands it as best represented as a continuum that has extremely formulaic expressions at one end (such as idioms and proverbs), and extremely novel expressions requiring full online processing and development at the other[6] (this is not the only way of viewing the notion of formulaicity; see Wray 2002; Chapter 3 for others).

Four criteria will be presented here as evidence of the formulaic tendencies of demonstrative clefts, namely, (1) their association with a specific type of text (namely, informal conversation), (2) their fixedness, (3) their fluent phonological structure, and (4) their non-salient reference. These properties are each discussed in the following sections, from section 3.1 through to section 3.4, respectively.

Before presenting evidence of the formulaic tendencies of demonstrative clefts, it is worth noting that to my knowledge, there is no work investigating the formulaicity of any cleft type (be they *it*-clefts or *wh*-clefts). This could be due to the fact that other cleft types are indeed not formulaic – an aspect which would further validate the separate treatment of demonstrative clefts from reversed *wh*-clefts. Or alternatively, it could be that no one has noticed formulaic aspects in other cleft types and perhaps this research will serve to bring awareness to the question of whether other clefts structures may be formulaic or not. The issue remains open for debate and no claims are made here about formulaicity of other cleft types.

3.1 Associated with informal conversation

As mentioned earlier, this paper reports on a study conducted by investigating approximately 200,000 words of spontaneous unplanned conversation extracts

6. It is interesting to note that the idea of "novel language" is engrained and assumed in our discussions of language to such extent that there are no terms to describe degrees or types of novelty, as there are with formulaic language; novel language has been (up to now) the norm against which we note the exceptional cases where fixedness of some kind or other comes in.

from the WSC. The data was manually coded for the various types of clefts found, including demonstrative clefts, which constituted the focus of the work. Findings show that overwhelmingly, demonstrative clefts are significantly[7] the most frequent type, as given in Table 1.

Table 1. Cleft frequencies in circa 200,000 words of conversation

Cleft type	Raw counts	Percentages
Demonstrative clefts	205	47%
It-clefts	145	33%
Wh-clefts	73	17%
Reversed *wh*-clefts	12	3%
Total	435	100%

These results support those of Biber et al., who note the high occurrence of the demonstrative cleft in conversation, a fact which they attribute to the construction's informality and repetitiveness (1999: 962–963).

As claims of the association between demonstrative clefts and spoken language, and in particular, informal conversation mount up, it must be said that since no other studies have looked at the demonstrative cleft specifically, it is difficult at this stage to know for sure that the construction is associated with these genres alone. It may be the case that the demonstrative cleft is (almost) exclusively found in informal conversation and no other forms of speech; or it may be that it is used in highly informal, personal and emotive written extracts too. More research is needed to confirm or disconfirm this. What is clear, however, is that there is indeed a link between the demonstrative cleft and informal conversation.

In light of the earlier discussion in Section 2, it is perhaps not surprising to find that the demonstrative cleft plays an important role in, and has tight connection to, informal conversation. As mentioned there, its role in managing and organizing the discourse through deictic links with the surrounding co-text, as well as low informative content make it an ideal tool for this linguistic genre.

According to Wray, formulaic constructions are used for three main purposes: in order to help the speaker by easing processing load (buying processing time, creating shorter processing route or manipulating information), in order to help

7. A two-tailed Chi Squared test shows very strong evidence against non-randomness of the distribution given in the table; x^2 (3) = 210.000, p > 0.0001. I am grateful to Alison Wray for pointing out one drawback of using the Chi Squared test, namely that we are forced to assume equal distributions of the various expected counts of cleft types, which may not necessarily be realistic given their differing functions and structures. Nevertheless, the method still provides reassurance that the effect we are seeing is real (in some way).

the hearer by making intentions and relationships clear (asserting group identity, asserting individuality or manipulating the hearer's world), and organizing and signaling the structure of the discourse (Wray 2000: 478 and Wray 2002: 97). It is in this third category that the demonstrative cleft can be placed. Two examples are given below to exemplify this role.

The first example illustrates how the speaker uses a demonstrative cleft to build up to a climactic point of the story (i.e., the point when she falls ill) and at the same time, to signal to the hearer that there is more to come and therefore, that she should not be interrupted (in other words, the speaker is securing the floor).

(8) FG: last time we were there ⟨laughs⟩
MJ: ringing london every day saying I'm not going to make it today
→ FG: we were booked to spend two nights in Amsterdam and **that's when I fell ill**
RW: oh
FG: and I kept thinking I would get better
RW: mm
FG: and that I would ⟨,⟩ be ⟨,⟩ okay to travel the next day so Glen would ring sand say well we won't be here tomo today but we will tomorrow
(WSC, DPC181:1610-1640)

The second example shows a demonstrative cleft which is used to link two ideas already mentioned in the discourse. The first point is that a previously mentioned female is gossipy, then the speaker informs the hearer that Bill is also gossipy. The demonstrative cleft is used to make a link between the two statements, i.e., the reason Bill is gossipy is because he takes after her (and she is gossipy).

(9) BG: oh no she is lovely SHE'S GOSSIPY THOUGH
AT: mm
→ BG: very gossipy LIKE BILL **that's where Bill gets it from**
AT: ⟨unclear word⟩ oh he is a little gossip talking about Mike Furley
(WSC, DPC096:1765-1790)

3.2 Fixedness

We now turn our attention to the structure of the demonstrative cleft. As given in (6), the construction involves a demonstrative pronoun, followed by a copula verb, a wh-word and a relative clause. The very fact that we have a cleft "formula" is suggestive of the presence of a frame with slots which are filled by various elements (something which also applies to the other cleft types in equal manner). It turns out, moreover, that even when it comes to the various elements allowed to fill these slots in demonstrative clefts, not all possibilities are realized.

First, there is the potential to add components (i.e., phrases) to the formula given in (6), as exemplified in (10) below.

(10) a. That is **exactly/potentially/perhaps** why he went home last night, after the game.
b. That is why, **I believe/I think/I would say**, James would have sold his car.

In (a), the element inserted is an adverbial phrase, functioning as an intensifier and describing the speaker's stance to the material expressed in the cleft clause. In (b), the cleft contains the variant epistemic phrases *I believe/I think/I would say*, which act as hedges, expressing the speaker's hesitation towards the assertion made (see Thompson & Mulac 1991 for details on epistemic phrases).

However, while these possibilities are theoretically available, they are only rarely seen in the WSC data. From the total of 205 constructions, only 3 (2%) demonstrative clefts contain epistemic phrases, and only 15 (7%) involve adverbial modifiers, illustrated in (11) and (12), respectively.

(11) FG: mm when ⟨,⟩ thumb sucking must stop if it hasn't before
RW: ⟨drawls⟩ well if it hasn't before yes and always
FG: mm or any anything that can be seen in ANY way
→ RW: as being childish yes yes **that's when** I THINK **these things get dropped** ⟨inhales⟩ but syntactically as well I hear all sorts of strange things from keith's friends which suggest to me that they really still at the age of ten don't have ⟨,,⟩ er ⟨,⟩ proper command of their ⟨,⟩ own language
(WSC, DPC182:0830-0855)

(12) FD: so anyway that night he used to watch the television a bit with me at half past six watch the news so I said to him what're you going to do you going to a flat or you going to ⟨,⟩ boar somewhere he said a very small flat he said ⟨6: 00⟩ that was all no information whatever but Sam told me ⟨,⟩ that he had been talking to him and he told him it was in Durham street this flat and said there's some awful places up there ⟨laughs⟩ he said ⟨/laughs⟩ but **that was where he was going to go** APPARENTLY
MG: yeah oh no yeah ⟨quietly⟩ oh okay ⟨/quietly⟩ ⟨laughs⟩
(WSC, DPC033:0570-0610)

Secondly, even within the represented slots, not all possibilities are realized. First, we consider the cleft constituent slot. Cleft constituents never involve plural demonstrative pronouns.

(13) * **These/those** are what/why/when/where/how you went home that day.

What is interesting is that even as far as the singular forms are concerned, the use of the proximal pronoun is very rare: only 13 (6%) examples were cited. The majority (94%) of demonstrative clefts contain the distal pronoun *that*.

One problematic aspect of *this*-demonstrative clefts is their reference in discourse. Previous literature suggests that discourse deictic *this*-constructions are cataphoric, pointing to upcoming discourse. For example, Fillmore writes: "the forward-pointing demonstratives of discourse deixis are similarly distinguished, I think, because when I say (just before giving my explanation) "This is my explanation", I know what it is but you don't; but when I say "That was my explanation," we both know what it is" (1997: 105). This claim is supported by the work of Miller (1996) and Miller and Weinert (1998). Diessel (1999) also mentions the use of demonstrative pronouns as discourse deictics, that is, as markers used to refer to propositions or speech acts. According to him, discourse *this* can involve both anaphoric and cataphoric reference. The constraint is that it can only be used to refer to a speaker's own utterances, not to the contribution of other participants (1999: 102–103).

However, contrary to claims stating the necessary anaphoricity of *this*-clefts, the examples of *this*-clefts found in the WSC data can indeed be anaphoric, as given in (14). The clefts uttered by speaker RW, *this is obviously going to be what they're what they should be doing at the end of it* and *this is what we have to take as a goal,* point to the previous utterance *the interesting question i mean as a sort of measure of this you obviously have to have body of adult speech to compare it with* (even though speaker FG repeats the value referenced by the second cleft, *the adult speech*, for clarification and perhaps to show that she is following the conversation).

(14) RW: yeah that I think is THE INTERESTING QUESTION I MEAN AS A SORT OF ⟨,,⟩
MEASURE OF THIS YOU OBVIOUSLY HAVE TO HAVE ⟨unclear word⟩ BODY
→ OF ADULT SPEECH TO COMPARE IT WITH AND SAY OKAY **this is obviously going to be what they're what they should be doing at** ⟨,⟩ **at the end** →**of it this is what we have to take as as the goal** ⟨latch⟩
FG: the adult speech mm yes ⟨latch⟩
RW: at what point do they actually reach that goal ⟨latch⟩
FG: because if the parents have ⟨latch⟩
RW: mm nonstandard or yes
FG: less than perfect command (WSC, DPC182:0880-0925)

Unfortunately, the small data size of only 13 examples prevents us from having a conclusive explanation for the role of *this*-clefts in discourse, and the issue is left for future work to determine.

We now turn our attention to the next element in the demonstrative cleft frame, the copula verb. The data from the WSC shows that the copula verb found in demonstrative clefts is never used in tenses other than the simple present and the simple past. Most commonly, the copula is used in the contracted *'s* form. Furthermore, it never occurs with any aspectual marking (e.g., ? *That has been what*

he thought), and only rarely with negative polarity (only 3 out of 205 clefts); see the use of negation in example (15). Finally, there is only one example where the copula is accompanied by a modal verb, as given in (16).

(15) MA: yeah well I think there's no you can't bloody claim that at all I mean that's ridiculous
FA: no I I disagree completely ⟨latch⟩
MA: why ⟨latch⟩
FA: why not I mean inherent in in the treaty is the right of development that that's been granted to the crown as well so why shouldn't maori have it
→ MA: yeah yeah no I'm **that's** NOT **what I mean** that's the thing there's a balance there ⟨„⟩ that you ⟨unclear word⟩ too busy are talking past each other
(WSC, DPC179:0460-0505)

(16) AR: what age was he
BT: he was it would be about fourteen or fifteen and THEN I said to HIM oh that's so strange because a few days ago I saw the same thing now several years later I picked up one of my books er cos I've got a lot of um nonfiction books and this particular one I just flicked it over and there it
→ said a a a ball of energy an orange light and I thought OH **that** MIGHT **be what happened** I must read that and I put the book BACK at that stage and I was so busy doing other things at the time and I went to look for it a few months later and I looked in every book in the bookshelves and I can't find it but I definitely saw that heading so I ⟨9: 00⟩ think and I have heard VAGUE comments of other people that there is a sort of an orange light an energy ⟨latch⟩
AR: mm ⟨latch⟩ (WSC, DPC121:0455-0480)

The next element in the demonstrative cleft frame is the wh-word. Most demonstrative clefts involve the use of *what*, *why* or *where*, and only rarely *when* and *how much*, as summarized in Table 2 (see previous examples of demonstrative clefts given throughout the paper).

Table 2. Wh-words and their frequencies in demonstrative clefts

Wh-word	Raw counts	Percentages
What	123	60%
Why	39	19%
Where	24	12%
How much	12	6%
When	7	3%
Who	–	–
How	–	–
Total	205	100%

Finally, we have the cleft (relative) clause. Even here, we find a certain degree of predictability. The subject of the cleft clause is in 90 per cent of cases highly given, either coded by a personal pronoun or a proper name. Furthermore, the predicate typically contains a verb of cognition (such as *think, believe, understand, know, wonder*), communication (e.g., *say, tell, ask, mean, talk, call*), or movement (*go, come, do, make*), with over half of the verbs occurring the first two semantic groups (cognition and communication). Examples (17) and (18) illustrate these two possibilities, respectively.

(17) CH: flashing is a bit of tin that they put between the roof and the thing like flashing on windows you shove a bit of tin up the top
DN: oh yeah right
BT: what to stop it ⟨,⟩ as a ⟨unclear word⟩
CH: so
AL: it's in
→ CH: yeah to stop it yeah **that's what** I THINK it is yeah but if it's not that then the drains are blocked and we'll have to get down there with a bit of wire
AL: between the ⟨unclear word⟩ to bring it down between ⟨13: 00⟩
DN: oh look I just can't stand chaps coming to my place to do jobs ⟨,⟩ I inevitably get that ⟨,⟩ sort of crap (WSC, DPC066:1620-1670)

(18) BH: the brace helps to hold you upright ⟨,,,⟩
UV: the only thing for a sore back is bed rest
→ BH: well **that's what** THEY SAY **eh**
UV: yep
BH: and heat (WSC, DPC214:0875-0895)

In light of these findings, over 90 per cent of all demonstrative clefts in the corpus can be captured by the formula given in Figure 1. The cleft constituent is predominantly the singular distal demonstrative pronoun, followed by a contracted form of the copula (in either simple present tense or simple past tense), a wh-word likely to be *what*, *why* or *where*, and a relative clause containing a personal pronoun or proper name in subject position and a verb of cognition, communication or movement as part of the predicate.

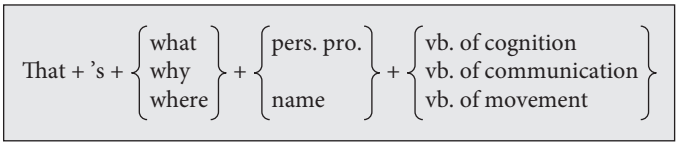


Figure 1. The demonstrative cleft formula.

3.3 Fluent phonological structure

A third signal of formulaicity in demonstrative clefts has to do with their phonological structure, or more precisely with their signs of "phonological cohesiveness" (Hickey 1993: 32–35). It has been observed that compared to novel expressions, formulaic ones exhibit higher overall fluency, are pronounced faster, and tend to have fewer pitch changes and fewer changes in intonation patterns (Wray 2002: 35). In other words, formulaic expressions are treated like individual words, not just in terms of their storage, but also in their delivery.

Demonstrative clefts show such phonological cohesiveness in their lack of pauses. Only six cleft examples (from a total of 205, i.e., 3%) contained any pauses during the course of the cleft, as exemplified in (19). Speaker TS is still thinking about the utterance she is about to produce when beginning the cleft *that's what I worry about is um eating to m*, and thus never quite manages to carry it out completely, changing her mind sometime during the course of uttering the cleft clause (*eating too m*). The pause precedes the cleft clause, occurring together with a discourse marker signaling hesitation (*um*).

(19) LU: yeah we c we could kill you with over kill you with eating too much chocolate ⟨,⟩ overdose on chocolate ⟨latch⟩
TS: oh no po poisoned ⟨,⟩ oh you mean just if there was just too much chocolate
KA: put a box of them out there ⟨latch⟩
→ TS: actually **that's what I worry about is um ⟨,⟩ eating too m** no is I eat enormous quantities of chocolate every day I'd have chocolate ⟨latch⟩
(WSC, DPC024:0395-0425)

Typically, however, if there are any pauses in close proximity to the demonstrative cleft, these occur after the cleft construction. In other words, they do not interfere with the boundary of the cleft, as it were, but similarly to the hypothesis formulated to explain observed practices from code-switching (Backus 1999), the formulaic expression is not "interrupted" by any changes (be they in terms of the language used, or the phonological structure adopted).

It is interesting to note that the pauses occur *after*, and never *before* the cleft construction, as in (20). The motivation for this pattern is unclear at present.

(20) LU: ⟨laughs⟩ SKINNY
TS: stupid word ⟨latch⟩
LU: it's going a bit far isn't it
TS: yeah
KA: yeah I'm skinny now eh ⟨latch⟩
LU: you're lean Ginny said you were lean Kay ⟨latch⟩
KA: oh that's nice can't have been me
→ TS: hey **that's what I was um ⟨„⟩** we listened to Dale Spender eh talk on the radio as well did y did you see her in Auckland (WSC, DPC024:0700-0750)

What does, however, seem to often proceed demonstrative clefts instead of pauses are discourse markers, another way of signaling boundaries, as given in examples (21) and (22). A quarter of all demonstrative clefts found in the WSC data (51 of the 205) contain discourse markers in this position (compared to only 14 cases occurring inside the cleft construction, and 14 following it).

(21) KK: oh that's interesting because I mean I'm not taking any raincoat or anything like that and just a sweatshirt and the rest is just
AN: want some chips mm want some chips mm you might need an umbrella
→ KK: well I'll buy one over there ⟨laughs⟩ YEAH **that's what I thought** it'll be very much like Auckland really muggy but raining so you can't wear a coat cos it's too hot but you need an umbrella um presumably it would be like that
(WSC DPC008,1095-1125)

(22) GG: so that Ricky Stewart man he should be playing in the bloody team
PT: mm
AS: turns on the mathematicians probably
GG: oh see after that the guy he scores that try the guy punched him in the face ⟨7: 00⟩
PT: ⟨laughs⟩ ⟨quietly⟩yeah⟨/quietly⟩
→ AS: OH WELL **that's what I'd do if a guy scored a try**
PT: ⟨laughs⟩
GG: ⟨laughs⟩ it's a HOOD'S game though (WSC, DPC 030:0880-0915)

3.4 Non-salient reference

The final piece of evidence of the formulaic tendencies of demonstrative clefts comes from their lack of transparency with regard to their reference in discourse, or their "non-salient reference", as termed by Hudson (1998).

In her 1998 doctoral dissertation, Hudson discusses the referential ambiguity of various expressions (demonstrative pronouns being included among these). She refers to this ambiguity as reduced salience or "non-salience" (see Chapter 7). Hudson observes that it is often not possible to tell what expressions such as *and all that*, *that's all*, and *that* refer to in a given text. Her arguments support psycholinguistics studies (such as Gibbs 1994), which show a correlation between reduced salience and fixedness in idioms. What is new in Hudson's approach is that she proposes a cline between more or less explicit reference (1998: 109), as given in Figure 2.

The cline predicts that the higher the level of salience a given expression will exhibit, the more fixed the expression (i.e., expressions at Level 3 are less variable and more formulaic than those at Levels 1 and 2). In other words, "maximum fixedness correlates with minimum salience" (Hudson 1998: 107).

As discussed throughout the paper, demonstrative clefts have tight deictic links with the discourse, through the use of demonstrative pronouns in cleft constituent

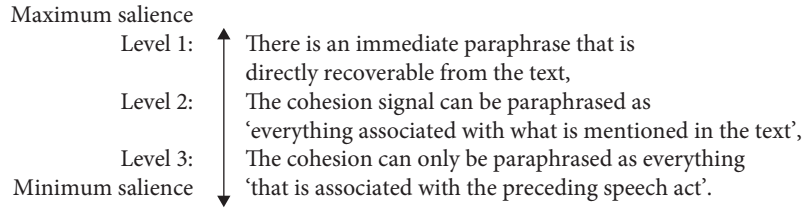


Figure 2. Hudson's referential continuum.

position. The data in the WSC shows that in some cases it is difficult to pinpoint exactly what the material referenced by the demonstrative cleft actually is. In other words, a cleft like *that's what I'm talking about* may be clear enough in giving a sufficient idea of the speaker's intention and opinion (one of agreement and appreciation perhaps), but at the same time, it may be difficult to pick out the exact referent of *that,* and indeed, to give an appropriate paraphrase of it. This point is illustrated with two examples below. Consider the demonstrative cleft in (23).

(23) JM: I went from Karori West to Wimbledon in Southern Hawkes Bay which was at that time ⟨drawls⟩ a very small sole charge school and I was the only child who was NOT related to anyone else
WC: that must have been quite a shock
XX: ⟨unclear word⟩
JM: it certainly was
XX: mm
WC: and how did you go about adjusting to that was it easy
JM: it wasn't VEry easy I can remember quite vividly being locked in the tool shed at ⟨drawls⟩play play times as a sort of ⟨,⟩ um ⟨,⟩ initiation ceremony
WC: BY the other kids?
JM: yes ⟨,⟩ by all the other kids
VV: goodness me
WC: that's charming isn't it
JM: ⟨drawls⟩ so it was charming but you know we finally got out of that ⟨latch⟩
→ VV: **that's what happens** children gang up when they have a stranger come in
JM: joined in (WSC, DPC060:0020-0100)

The cleft construction *that's what happens* appears to point to the entire preceding material detailing the "initiation ceremony" which new children are subjected to by their peers upon joining a new school. While this much is clear, it is not actually possible to express in a complete paraphrase exactly what *that's what happens* really refers to. This is because the cleft does not point only to the incident of being locked up in the tool shed, but to a much bigger set of events, namely all

those nasty experiences which pupils (in the "out" group) experience, in general, when being bullied by the locals (the ones in the "in" group). This means that the demonstrative cleft in (23) would be located at Level 2 on the referential continuum proposed by Hudson (1998).

Another example of reduced referential salience is given in (24). Here, once again, it is impossible to offer an appropriate paraphrase of *what it is all about*. The cleft presumably refers to the entire relationship of being friends and helping out.

(24) AS: Megan's got one
PP: ⟨unclear word⟩
AS: or a sleeping bag
PP: yeah
XX: thank you very much sorry to you know ⟨laughs⟩ come and crash at short notice ⟨/laughs⟩
→ AS: well **that's what that's what it's all about**
XX: Megan's always good to me like this
AS: yeah so how long have you got to go in your course
(WSC, DPC078:01870-0910)

These examples above show a reduced referential salience of (some) demonstrative clefts, which under the interpretation proposed by Hudson, is indicative of an increased fixedness in the construction.

4. Summary

As already pointed out by a vast number of studies, spoken language contains a high number of routines, formulae and fixed expressions. Among these is the most frequent cleft type found in spontaneous, spoken language, namely the demonstrative cleft. It exhibits a number of formulaic properties, such as structural fixedness, phonological fluency, and reduced referential salience. As shown in the present study, the vast majority (90%) of demonstrative clefts found in conversational English can be captured by the formula [*that +'s/was* + *what/why/where* + [personal pronoun/proper name + verb of cognition, communication or movement]]. While demonstrative clefts show tendencies of formulaicity, it remains to be shown exactly where their place is on the formulaic ⇔ novel continuum. Furthermore, this paper raises questions regarding the formulaic status of other cleft types (i.e., *it*-clefts, *wh*-clefts and so on). Finally, the analysis of the demonstrative cleft contributes to our understanding of the complex and controversial notion of formulaicity.

Appendix A: Discourse features included in the examples

⟨drawls⟩	speaker drawls
⟨drawls⟩ … ⟨/drawls⟩	speaker drawls for the duration of the utterance as given inside the markers
⟨inhales⟩	speaker inhales
⟨latch⟩	overlapping speech
⟨laughs⟩	speaker laughs
⟨laughs⟩ … ⟨/laughs⟩	speaker laughs for the duration of the utterance given inside the markers
⟨quietly⟩ … ⟨/quietly⟩	speaker utters the material inside the markers quietly
⟨sighs⟩	speaker sighs
⟨unclear word⟩	unclear word
⟨,⟩	1 second pause
⟨,,⟩	2 second pause
⟨,,,⟩	3 second pause
⟨7: 00⟩	indicates a certain number of second pause, whatever the number indicated in the ⟨ ⟩ brackets, e.g., ⟨7: 00⟩ corresponds to a 7 second pause
?	signals an interrogative, where it would be ambiguous on paper otherwise (between a statement and a question)

Acknowledgements

I am indebted to the organizers of the Symposium for Formulaic Language, at UWM, in Milwaukee, Wisconsin (April, 2007) for the stimulating conference they put together and hosted. Additionally, I am thankful to the conference participants and their feedback to my talk on demonstrative clefts. In particular, I would like to thank Alison Wray for her comments (especially relating to the statistical tests used), Jean Hudson, for her discussion regarding the non-salient reference of demonstrative clefts (and for kindly sending me her thesis), and Sandra Thompson, for her interest in and discussion of demonstrative clefts. Finally, the paper was much improved by the meticulous reading of Frank Lichtenberk, to whom I am, as always, much obliged, and by the comments and suggestions made by the referees and book editors.

References

Aijmer, Karin. 1996. *Conversational routines in English: Convention and creativity.* London: Longman.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In *Phraseology: Theory, analysis and applications,* A. Cowie (Ed.), 101–122. Oxford: Clarendon Press.

Backus, Ad. 1999. Evidence for lexical chunks in insertional codeswitching. In *Language encounters across time and space,* E.L. Brendemoen & E. Ryen (Eds), 93–109. Oslo: Novus Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Essex: Longman.

Biber, Douglas & Susan Conrad. 1999. Lexcial bundles in conversation and academic prose. In *Out of corpora: Studies in honour of Stig Johansson,* H. Hasselgard and S. Oksefjell (Eds), 181–190. Amsterdam: Rodopi.

Calude, Andreea. 2008. Demonstrative clefts and double cleft constructions in spontaneous, spoken English. *Studia Linguistica* 62(1): 78–118.

Collins, Peter. 1991. *Cleft and pseudo-cleft constructions in English*. London: Routledge.

Collins, Peter. 2004. Reversed *what*-clefts in English: Information structure and discourse function. *Australian Review of Applied Linguistics* 27(2): 63–74.

Diessel, Holger. 1999. *Demonstratives: Form, function, and grammaticalization*. Amsterdam: John Benjamins.

Fillmore, Charles. 1997. *Lectures on deixis,* 2nd Edn. Stanford CA: CSLI.

Ford, Cecilia, Barbara Fox & Sandra Thompson. 2002. Social interaction and grammar. In *The new psychology of language,* Vol. 2, M. Tomasello (Ed.), 119–143. Mahwah NJ: Lawrence Erlbaum Associates.

Gibbs, Raymond. W. 1994. *The poetics of the mind.* Cambridge: CUP.

Hedberg, Nancy. 2000. The referential status of clefts. *Language* 76(4): 891–920.

Herriman, J. 2004. Identifying relations: The semantic functions of wh-clefts in English. *Text* 24(4): 447–469.

Hickey, Tina. 1993. Identifying formulas in first language acquisition. *Journal of Child Language* 20: 27–41.

Holmes, Janet., Bernadette Vine & Gary Johnson. 1998. *Guide to the Wellington corpus of spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Hopper, Paul. 1987. *Emergent grammar* [Berkeley Linguistics Conference 13: 139–157].

Hopper, Paul. 2001. Grammatical constructions and their discourse origins: Prototype or family resemblance? In *Applied cognitive linguistics*, Vol. 1: *Theory and language acquisition*, M. Pütz, S. Niemeier & R. Dirven (Eds), 109–129. Berlin: Mouton.

Huddleston, Rodney D. & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language.* Cambridge: CUP.

Hudson, Jane. 1998. Perspectives on fixedness: Applied and theoretical. Ph.D. dissertation, Lund University.

Lambrecht, Knud. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3): 463–516.

Miller, Jim. 1994. Speech and writing. In *The encyclopedia of language and linguistics, R.E. Asher & J.M. Y. Simpson (Eds), 4301–4306.* Oxford: Pergamon.

Miller, Jim. 1996. Clefts, particles and word order in languages of Europe. *Language Sciences* 18(1–2): 111–125.

Miller, Jim & Regina Weinert. 1998. *Spontaneous spoken language: Syntax and discourse*. Oxford: Claredon.

Oberlander, J. & J. Delin. 1996. The function and interpretation of reverse wh-clefts in spoken discourse. *Language and Speech* 39(2–3): 185–227.

Read, John & Paul Nation. 2004. Measurement of formulaic sequences. In *Formulaic sequences* [Language learning and language teaching 9], N. Schmitt (Ed.), 23–35. Amsterdam: John Benjamins.

Thompson, Sandra A. 2002. 'Object complements' and conversation: Towards a realistic account. *Studies in Language* 26(1): 125–163.

Thompson, Sandra A. & Anthony Mulac. 1991. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In *Approaches to grammaticalization,* Vol. II, E.C. Traugott & B. Heine (Eds), 313–329. Amsterdam: John Benjamins.

Weinert, Regina & Jim Miller. 1996. Cleft constructions in spoken language. *Journal of Pragmatics* 25: 173–206.

Wray, Alison. 2000. Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics* 21(4): 463–489.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.

Wray, Alison. 2006. Formulaic language. In *Encyclopedia of language and linguistics,* K. Brown (Ed.), 590–597. Cambridge: Elsevier.

Formulaic language and the relater category – the case of *about*

Jean Hudson & Maria Wiktorsson
Malmö University, Sweden

Abstract

Relaters in English are words that are traditionally referred to as prepositions, particles, or adverbs, such as: *in, for, with, at, about.* They are highly decategorialized and polysemous. Furthermore, with few word forms which occur very frequently, they are problematic for linguists and learners alike. In our data (naturally occurring conversation in the BNC), we find many formulaic patterns around these words, providing new insights into the meanings created through and around them. In this paper we report on a study of the relater *about*. We find that the greater proportion of the occurrences can be adequately described as part of substantive or schematic constructions, and that they to a large extent pattern with meanings with a negative or generally unfavourable orientation.

1. Relaters

A persistent problem for linguists and for learners of English is the relatively small group of words which, in the current descriptive tradition, are analyzed alternately as preposition, particle, adverb, conjunction or adjective. Some of the most frequent

of these are: *in, of, to, for, on, with, at, like, about, from, by, into, over, after, off, through, around, as, past, up, down, till, without, under, before*. They are not only difficult to classify grammatically, but also highly polysemous, as many researchers have demonstrated (e.g., Brugman 1988; Tyler & Evans 2003). At the same time they are extremely frequent: using a number of different corpora, in searches on the 40 most frequent word forms, we found that one of them turns up in approximately every 10th word. We call these words *relaters*.

In 1991, John Sinclair reported that *of,* the most frequently occurring relater, no longer behaves like a preposition. "Only occasionally, and in specific collocations with, for example*, remind*, does it perform a prepositional role. Normally it enables a noun group to extend its pre-head structure, or provides a second head word." (Sinclair 1991; republished in Sinclair & Carter 2004: 18). Examples of formulaic sequences with *of* are highly frequent patterns like: *a lot of work*, *a glass of milk*. Such careful and open-minded study of grammatical items in large corpora helps us to see patterns and structures that our previous 'knowledge' of linguistic categories keeps us from seeing otherwise. The same passage continues: "In due course the grammatical words of the language will be thoroughly studied, and a new organizational picture is likely to emerge." Through an ongoing research project on the relaters, we hope to make a small contribution to the attainment of his goal.

The aim of the present paper is to present some exploratory research that we have carried out on the relater *about* in order to investigate the potential for describing the relaters in terms of the formulaic sequences that surround them. A further aim is to thus contribute to the discussion on the nature of formulaicity, and the search for an adequate and useful model of description based on an understanding of the formulaic nature of English.

As to existing theoretical frameworks, we have found Construction Grammar (Fillmore, Kay & O'Connor, 1988; Goldberg, 1995; Croft & Cruse, 2004) to be very useful. It is important to note, however, that we avail ourselves of insights from other directions within the general framework of cognitive and functional linguistics. We begin with a, necessarily, brief account of some theoretical considerations that we find relevant to the discussion of formulaicity.

2. Formulaicity and formulaic language

Formulaicity is an all-pervasive property of meaningful linguistic expression. In a sense, therefore, the expression 'formulaic language' is an anomaly, since all language is at some level and to some extent formulaic. Some support for this view comes from work on emergence in language:

> The memory representation of language consists of units that can constitute utterances or intonation units, that is, not just words, but also *phrases and constructions*. The smaller units familiar from structural analysis – stem morphemes, grammatical morphemes – are not independent units, but rather emerge from these larger stored units via a network of connections among them. (Bybee 1998: 432, emphasis added.)

It seems only logical to assume that the network of relations that emerge between smaller units within and beyond the larger units is evidence of the inherent patterning that is a crucial part of a speaker's 'knowledge' of their language. While it is true that some elements are potentially infinitely variable, such as X in [*a book entitled X*], it remains that for the most part we negotiate meaning through patterns in discourse:

> The notion of Emergent Grammar is meant to suggest that structure, or regularity, comes out of discourse and is shaped by discourse in an ongoing process. Grammar is, in this view, simply the name for certain categories of observed repetitions in discourse. […] Its forms are not fixed templates but emerge out of face-to-face interaction in ways that reflect the individual speaker's past experience of these forms, and their assessment of the present context, including especially their interlocutors, whose experiences and assessments may be quite different. (Hopper, 1998: 156)

How, then, are we to interpret the expression 'formulaic language', which was probably the most frequently used expression at the Milwaukee symposium, albeit among interlocutors "whose experiences and assessments [were] quite different"?

Language lives in discourse and is constantly and continuously affected by speakers' past experience and present needs. In the process, not only do patterns of smaller units emerge from larger patterns, but also smaller units have a tendency to pattern in a linear form. The process by which this comes about is no different to that described by Hopper in the citation above. Hudson (1998) describes this process thus:

> … ad hoc expressions[1] take on new meanings through pragmatic inferencing in the discourse.[2] The development proceeds through semantic and phonetic reduction[3] to a stage at which the contribution of the parts of the expression to the whole is beyond conceptualization, and the expression becomes fixed in its realization. At this third stage expressions are completely invariable although they might still comprise more than one orthographic word. (Hudson, 1998: 2)

---

1. In the account presented in the present paper "ad hoc expressions" are also formulaic patterns, or, in the CG account, 'schematic constructions' with no specification of lexical material.

2. Traugott (1982).

3. Brinton (1996: 52–54).

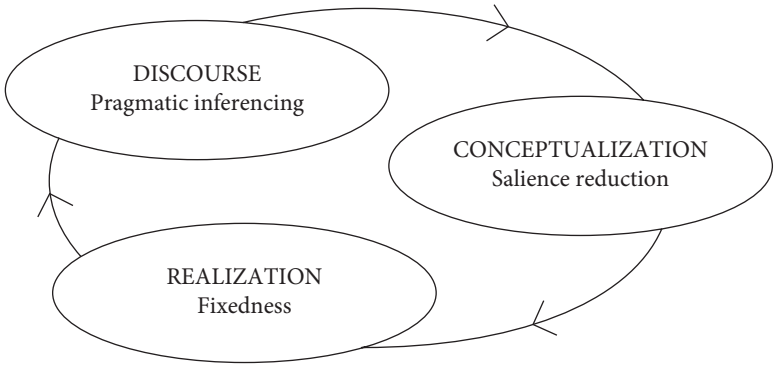


Figure 1. The process whereby expressions become fixed.

Space does not allow for a full account of this process, but by way of exemplification we refer to Traugott's (e.g., 1995) account of the development of the discourse markers *indeed, in fact*, and *besides* from regular, variable expressions functioning as clause internal adverbials, through pragmatic inferencing, to a state of univerbation, or the "welding of a syntagm into one word" (Lehmann, 1995: 151). At this stage, the meanings of the component parts are no longer salient in relation to the whole. This finding is relevant to our search for a working definition of 'formulaic language'. At this point, however, we need to say a few words on the notion of 'salience reduction'.

Briefly, salience reduction is the extent to which the behaviour of a word conforms to speakers' conceptions of how, based on past experience, the word usually behaves. Some examples of reduced salience can be found in the expression *all of a sudden,* where we find unusual features such as the use of *sudden* as a nominal and the highly decategorialized item *all*. Extreme decategoriality, like polysemy, is often concomitant with reduced salience in the sense that interlocutors must search for further contextual clues as to which meaning is intended in any given utterance. Vague or oblique reference is also a salience reducing feature. Compare (1) and (2):

(1) I wish I had eight lifetimes – it would take me all **that** to keep up[4]

(2) Finding comedy through character is only part of what Robin Williams does. […] Instantly Williams is off and running again, in the character of an unctuously anti-Semitic English headmaster: We're so happy to have you and all **that**, but Gawd, I'm sorry we don't have any of your food heah. What

4. From the London Lund Corpus [1.10.1188].

> is it that you people actually eat? And will you be doing any of your rituals while you're heah?[5]

In (1) the antecedent of *that* is quite salient (eight lifetimes). In (2) it is not possible to identify any specific referent or paraphrase; the only plausible explanation seems to be that *that* refers not to anything in the text or context (e.g., *we're so happy to see you*) but to the speech act GREETING, which would give the paraphrase: 'and all other things associated with GREETING'. This interpretation explains the pragmatic force encoded in *and all that* which we perceive, in the example given, as belittling or even dismissive, since the speaker has decided to abandon the appropriate and expected forms of greeting related to the situation.

In an investigation of expressions with the word *all,* Hudson (1998) finds that the extent to which salience reducing features are found in expressions is concomitant with the extent to which the expressions are fixed. In other words, if the sum of the most salient interpretations of the parts does not make sense in relation to the implied – and understood – meaning of the whole, the potential for variability of the parts is reduced to a similar degree. Thus, in the case of the invariable, or fixed, expression *and all that,* we find reduced salience both in the extremely decategorialized *all* and in the absence of an immediately identifiable referent for *that*.

Returning now to the process whereby expressions become fixed. More interesting, for present purposes, is not so much the final, univerbation stage of the process, but the very early stages of reanalysis where, through pragmatic inferencing in discourse, seemingly 'open', 'ad hoc' forms of expression nonetheless seem to be to a certain extent constrained. We suggest that it is somewhere around this stage in the development of more fixed, idiomatic forms of expression that the notion of 'formulaic language' becomes useful. What we (the analysts) then see, at the level of realization, is what Pawley & Syder, in 1983, so aptly described as "the puzzle of native-like selection".

In summary: observed repetitions of smaller units in larger ones, open up for analagous use elsewhere, and observed repetitions of certain combinations of smaller units in specific discourse environments lead (through reanalysis) to fixation and, at the extreme end, univerbation. Underlying both these processes is the attention to patterning that is an inherent component of our cognitive make-up. Thus, while **formulaicity** is a somewhat abstract property of language, **formulaic language** is, by our proposed definition, **any sequence of two or more words that are perceived to be more constrained than usual in their co-occurrence**. This somewhat vague definition is necessarily so: Since all language is formulaic, we are

5. Morgenstern, J. 'Stand up Robin Williams'. *The Guardian*. 5 January 1991.

dealing with a gradient category along a cline of less formulaic to more formulaic (as opposed to non-formulaic vs. formulaic).

The words in the relater category are both polysemous and decategorial, and, as we have shown above, polysemy and decategoriality are regular features of formulaic language. The hypothesis underlying the present paper (and the larger project) is thus that we will find much formulaic language around the relaters.

3. Construction Grammar

While we adopt an eclectic approach in our study of *about,* we conduct our analysis within the general framework of Construction Grammar (CG),[6] the ultimate goal of which is to capture all the regularities and irregularities of the language in a combined systematic description of form and meaning. A comprehensive description of the English language in CG terms remains a thing of the future, but studies carried out so far, such as Fillmore, Kay & O'Connor (1988) on *let alone,* Goldberg (1995) on argument structure, Kay & Fillmore (1999) on *what's that X doing Y*, and Tomasello (2003) on language acquisition, are interesting for all who work on formulaicity in language. Wiktorsson (2003), for example, investigates learners' use of idiomatic English by means of detailed analyses of native-like formulaic sequences used by the learners in their written production. She argues that CG provides a useful framework for the incorporation of these items in a description of language since it incorporates constructions at different levels of schematicity, from completely substantive constructions to the highly schematic.

According to (Croft & Cruse, 2004: 255),[7] (mostly) substantive constructions are, essentially, idioms, where all the lexical material is specified: [*kick-* TNS *the bucket*], while (mostly) schematic constructions are representations of what we have traditionally recognized as syntax: [SBJ *be-* TNS VERB *-en by* OBL]. In our analysis, we extend the extreme ends of this cline to completely substantive constructions [*by and large*] and completely schematic constructions [SBJ VERB], and we stress the gradient nature of the cline.

6. See Croft & Cruse (2004) for an overview.

7. We are not concerned with the 'complex'—'atomic' opposition, which distinguishes between expressions and words, since we are only concerned with expressions here.

4. Data and method

The data for the investigation come from the subset of the British National Corpus (BNC)[8] which comprises roughly 4.2 million words of conversation (s-conv). The decision to use only conversational data was based on the assumption that this is the language form that most people operate with for most of their time.[9] To search the data we used Mark Davies' (Brigham Young) interface at http://corpus.byu.edu.

In searching for form–meaning patterns (constructions), we must pay attention to semantic and pragmatic values in and around the patterns. The corpus that encodes such values is not yet available; we are therefore obliged to carry out a close reading of each and every concordance line. Given that the word form *about* occurs 13,105 times in the subcorpus s-conv, we decided to break down the material into datasets that are to a certain extent patterned in advance, that is, according to the grammar tag of the word preceding *about*. We then imposed one further limitation on the data by selecting only those occurrences of *about* with the sense of 'concerning'. Making an a priori distinction based on word sense might be considered a somewhat dubious approach, since categorization according to word sense is rarely an either/or affair. This is not so in the case of *about*, however, and we had no difficulty distinguishing the following three different senses of *about* in the data: 'concerning' (3), 'approximately' (4), and 'around' (5).

(3) It's a shame ABOUT Arthur though isn't it?

(4) mind you there was one year ABOUT six years ago

(5) Well you can't afford to mess ABOUT like you do!

The present paper is an account of the formulaicity around those occurrences of *about* that are preceded by a NOUN or an ADJECTIVE and where the sense of *about* is 'concerning'. We also add a brief comment on some initial findings on patterning around VERB+*about*. Eventually, the separate datasets of [X *about*] will be entered into one main database in order to facilitate cross sorting for similar features through the full corpus of *about* from spoken conversation in the BNC.

---

8. www.natcorp.ox.ac.uk

9. See also Hopper (1998: 171–172) for a discussion which comes out in support of the primacy of speech.

5. Results

Figure 2 shows the relative frequencies of different word classes that appear to the left of *about* (in all its senses). Verbs are most frequent; of the open classes, adverbs, nouns, and adjectives follow. The ADVERB+*about* group (n = 1374) is, surprisingly, less complex, and will be described in a separate report on the full dataset for *about*. Our focus here, as already mentioned, is on ADJ+*about* and N+*about*.
Of the 13,105 occurrences of *about* in this sub-corpus of 4.2 million words, almost half are immediately preceded by a VERB. Of the remainder, 616 are immediately preceded by an ADJECTIVE and 1,125 by a NOUN. After the usual filtering process, the results of which can be seen in Table 1, we were left with 538 for ADJ+*about* and 606 for N+*about* where *about* has its most frequent sense of 'concerning'.

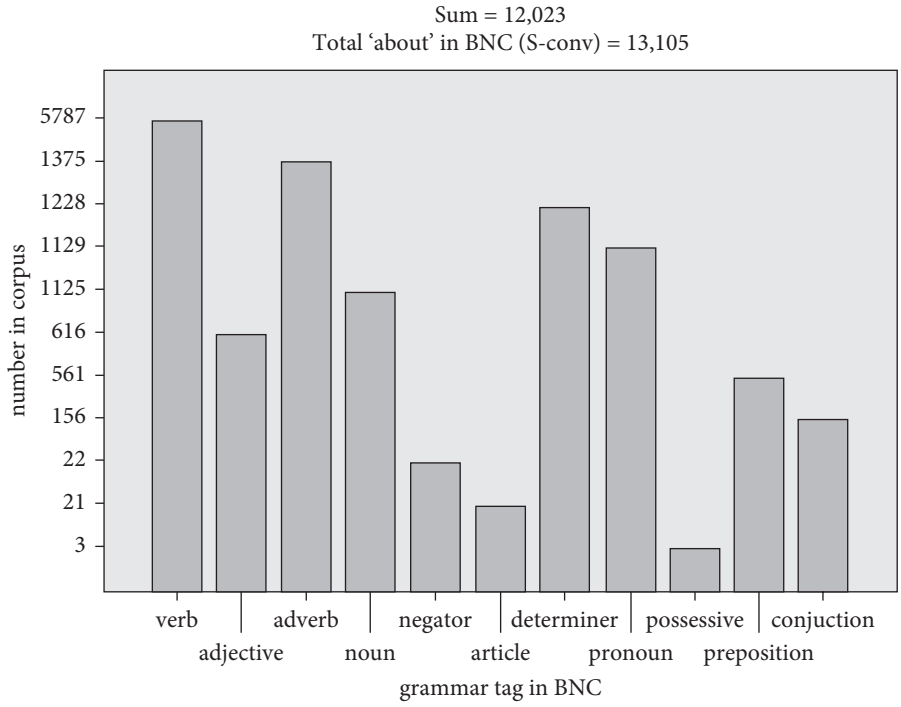


Figure 2. Breakdown of wordclass tags immediately preceding about in the BNC (s-conv).

What we can immediately see from Table 1 is that ADJ+*about* is much more likely to form a meaningful sequence (or part of one) than is N+*about*, where 10% of the hits have an intervening constituent, as in (6).

(6) there was this programme **last night** ABOUT the B B C when it first started in the nineteen fifties

Table 1. The [NOUN/ADJ *about* ('concerning')] datasets

	Noun + *about*		**Adjective + *about***	
Total hits in BNC s-conv	**1126**	%	**616**	%
Duplicates	6	1	9	1
Mistakes in tagging	76	7	29	5
about not 'concerning'	293	26	16	3
about not structurally related to the item immediately preceding	113	10	0	0
Unanalysable	32	3	24	4
Final database	**606**	54	**538**	87

Further, the sense of *about* is almost always 'concerning' in the ADJ+*about* pattern, which is not the case in the N+*about* set. Taken together, these two observations show that, despite the fact that *about* occurs immediately after a noun twice as often (1126) as it follows an adjective (616), the ADJ+*about* pattern is more often a meaningful sequence than the N+*about* pattern.

5.1 ADJECTIVE+*about*

The majority of adjectives that occur before *about* have a clearly negative semantics.[10] The more frequent of these are shown in Table 2 (with number of occurrences in parentheses).

Table 2. Adjectives occurring before *about* (frequency > 2)

worried	112	bothered	8	better	4	adamant	3
sorry	58	funny ('odd')	8	bitter	4	alright	3
sure	34	guilty	8	depressed	4	awful	3
concerned	22	nice	8	enthusiastic	4	embarrassed	3
happy	22	excited	7	funny	4	glad	3
pleased	20	right	6	paranoid	4	honest	3
upset	19	angry	5	rude	4	mad	3
good	11	annoyed	5	serious	4	weird	3
bad	9	fussy	5	uptight	4		

Some examples in context are shown in (7)–(9).

(7) we don't want want to make them feel awkward ABOUT it I mean they have planned what they can plan

10. For the pilot study we have relied solely on intuition in deciding what is and what is not a 'negative semantics'. Borderline cases were very few. In future studies this will be supplemented by informant testing.

(8) to go for it yes, but then what's bad ABOUT that is, the doctor's taking the supplies from the Health Service

(9) dear er in next bed to our Beryl She's ever so bitter ABOUT it then about fifty she's fifty two

Positive adjectives with *about* occur most often within the scope of a negator (10)–(12).

(10) when I spoke to Carole and she wasn't happy ABOUT it. Caro well Carole is problem, I don't think

(11) That's really nice that is! I'm not sure ABOUT the white one though. Let's go and watch that.

(12) thinks she's just left. They weren't too pleased ABOUT it, so they kicked her out the house I think.

Table 3 shows the extent to which ADJ+*about* correlates with a negative semantics and/or a negator in the data (actual occurrences, i.e., tokens).

Table 3. Polarity of adjectives preceding *about*

		SEMANTICS OF ADJECTIVE	
		negative	**not negative**
ADJECTIVE NEGATED?	negated	54	62
	not negated	305	117

Thus, in all of 421 cases out of 538 the word *about* is embedded in a negative environment.

Closer investigation of the (mere) 20% of cases where the adjective has a positive semantics and is not negated shows that there is nevertheless a tendency towards a more subtle expression of negativity. Frequently, there is in the broader context an indication that the ADJ+*about* sequence signals an improvement in an otherwise negative state of affairs, as in (13)–(16). Some of these occur in interrogatives, where there is a negative presupposition (17).

(13) Yeah. And, you know, I feel better ABOUT it now. You know, it made me feel a bit sick

(14) every time I go swimming Oh, you're very brave ABOUT it dear, er Margaret W one, that British Polio Fellowship

(15) Oh yes and I had er pressure sores, you know? Mm. Yeah, no I think you just need to be frank with him and you know say you're happy ABOUT the finger, fine, but just, you know, I'm more concerned about these Yeah.

(16) I don't know, he's just got a different approach about it, don't know much better. What? They're better? Well not necessarily better, they're just more mature ABOUT it. I'm sure when they're all sitting around in their studies they're a complete, they're all the same but But I suppose they're all boys really, I mean are all boys like this really?

(17) no no we don't want that rubbish what's interesting ABOUT that? ooh oh it's on till midnight yes it is

One conclusion that we draw concerning the pattern ADJ+*about* is that it is a schematic construction with a strong semantic prosody of negativity (Figure 3).

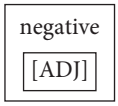


Figure 3. Representation of the [ADJ *about*] construction.

Looking beyond the [ADJ *about*] core, we find that the majority of the 538 occurrences are actually part of a somewhat more substantive construction: [*be/get/feel*-TNS (ADVL) ADJ *about* X], exemplified in (18)–(20).

(18) I was quite annoyed ABOUT that

(19) she didn't want you to be competitive ABOUT it

(20) he was very cagey ABOUT the fault on that …

Of course, this construction is to a certain extent predictable, since only predicative adjectives can be followed by a preposition and since the sense of *about* is always 'concerning' in the data. But we argue nonetheless that it is useful to identify such frequent patterns and to describe their formulaicity in constructional terms.

Only one completely substantive construction appears in the data, with 39 tokens: [*sorry about that], exemplified in (21)–(23).*

(21) you want a game of footy, I said oh sorry ABOUT that love, well I'll go and …

(22) True. Sorry ABOUT that. A slip of the finger. Naughty mummy. Sorry

(23) Mm. Sorry ABOUT that. Oh. You just couldn't tell from the outside

The remaining 19 tokens of *sorry about* are part of only slightly more schematic constructions, as in (24)–(26).

(24) poor old Arthur! I'm very, very, sorry ABOUT that!

(25) I'm sorry ABOUT it because you know, how I feel …

(26) Erm sorry ABOUT this

The substantive *sorry about that* construction has become a fixed, idiomatic expression which has clearly started off on a path towards univerbation. Symptomatic of this process are features such as reduction (*I'm sorry …* → *Sorry …*) (Hopper 1991); increased discourse function (routinized form of apology) (Traugott 1982); and vague reference (*that*) (Hudson 1998). The (smaller) number of more sche-

matic constructions exhibit slightly more variability, though common to all of the 58 occurrences of *sorry about* is the constraint on the subject position (whether or not it is explicitly pronounced) to first person singular.

5.2 NOUN+*about*

The pattern of formulaicity in the NOUN+*about* set is more complex than in the ADJ+*about* set, but nonetheless pervasive. Table 4 shows the nouns preceding *about* with a frequency of more than two.

Table 4. Nouns occurring before *about* (frequency > 2)

thing	94	shame	9	paper	5	business	3
things	24	idea	8	article	4	complaint	3
bit	20	song	8	conversation	4	deal	3
doubt	20	stories	8	details	4	joke	3
lot	17	film	7	discussion	4	morning	3
story	15	argument	6	fuss	4	noise	3
talk	15	comments	6	jokes	4	opinion	3
information	12	laugh	6	letters	4	programme	3
questions	11	point	6	part	4	row	3
news	10	question	6	problem	4	think	3
book	9	ado	5	trouble	4		
mind	9	ideas	5	worry	4		

The first thing to notice is that we do not see among these nouns the high degree of negative semantics that we found in the adjectives. What we do find is that we can arrange the dataset for NOUN+*about* as constructions, along a substantive—schematic cline (Table 5).

Table 5. Prototypical constructions in the N+*about* dataset. (A=substantive, B=borderline or uncertain, C=schematic)

SUBSTANTIVE

	Prototype constructions along the scale A-C	**Example**
A1	Completely substantive and fixed	Much ado ABOUT nothing
A2	[*have-* TNS/ASP/PERS (*got*) DET *thing about* X]	had this thing ABOUT X has got a thing ABOUT X
B1	Potentially substantive	What's all this noise ABOUT?
B2	[*it was in the paper/on telly/on the television/radio/news about* X]	it was in the paper ABOUT X
B3	[VERB VAGUEQUANT *about* X]	know a lot ABOUT X
C	[N *about]*	book ABOUT X

SCHEMATIC

5.2.1 *Substantive constructions (type A)*

There is only one completely substantive construction (A1): [*much ado about nothing]*. Type A2 we have classified as substantive since all the lexical material is fixed; the only variation is in grammatical elements. Approximately one third of the hits in the dataset (some 200 tokens)[11] are part of longer constructions that are largely substantive and occur frequently (27)–(31).

[*have-* TNS *a* (MODIFIER) *laugh about* X]

(27) So we all had a laugh ABOUT it when we went off

(28) Oh we had a right laugh ABOUT it

[*make-* TNS (DET) *a big thing about* X]

(29) they made a big thing ABOUT us sort of not using one millimetre holes

(30) they make a big thing ABOUT some things, and then some we're gonna …

(31) they shouldn't make such a big thing ABOUT all these skinny models

These more variable (with respect to tense, aspect, modality etc.) type 2 constructions generally contribute to a higher degree to the propositional content of the utterance than the less variable ones, which instead tend to have pragmatic functions, e.g., are attitudinal markers or discourse structuring items (32)–(37).

[*it's a shame about* X]

(32) Yeah. funny really. It's a shame ABOUT that. Weird!

(33) It's a shame ABOUT Arthur though isn't it?

(34) It's a shame ABOUT leaving Hounslow where Chiswick is …

[*no doubt about it/that*]

(35) Some people take advantage of au pairs, no doubt ABOUT it. I mean you remember I told you about my cousin …

(36) Well I'm doing my best! No doubt ABOUT it, we'll never get it finished!

[*thing about* X *is*]

(37) Yeah Manchester. Carl! What all those? Thing ABOUT up north is they got good woman [*sic*]

---

11. Because of the scalar nature of the substantive—schematic cline we see no sense in attempting to enumerate in detail the exemplars in each type group.

In (35)–(37) we see once more features of fixedness similar to those observed in (21)–(23) with *sorry about that,* namely: reduction, strengthening of discourse function, and general reference.

5.2.2 *Borderline constructions (type B)*

Some of the constructions are borderline between substantive and schematic. They are so for different reasons, thus the classification as B1, B2, and B3 does not indicate any kind of cline within the type B group.

In B1 we place constructions that we suspect are substantive but for which we have not (yet) found supporting evidence. The expression *What's all this noise about?* is reminiscent of the [*What's X doing Y?*] construction (Kay & Fillmore 1999) in that it has a similar feature 'undesirable' as a function of the whole expression but not of any of its parts. If it transpires that *noise* can be substituted by a number of other nouns, leaving all other semantic and syntactic features of the expression unchanged, then we have a slightly more schematic construction: [*What's all this* X *about*?]. This we do not yet know.

Type B2 comprises borderline cases proper in our N+*about* data. Here, the lexical material is only slightly variable within the more schematic construction [*it was in*/*on* X *about* Y].

In the B3 type there is no affinity between *about* and the preceding noun. These nouns are themselves involved in fixed expressions (*a lot*, *a bit*) but were automatically trawled up in the search for N+*about*. They would have been eliminated from the dataset but for the fact that a) they are numerous, and b) we suspect that they might contribute to a generalization when we conflate the datasets of each wordclass followed by *about*. We do not make any further comment on this type in the present paper, though they will reappear in the fuller account of constructions around *about*.

5.2.3 *Schematic constructions (type C)*

Approximately half of the occurrences of NOUN+*about* (some 300 tokens) are constructions with what at first sight appears to be a schematic slot for a noun followed by the substantive element *about*. They have a straightforward semantics, are grammatically variable, and do not seem to be part of any larger construction (38)–(40).

(38) It's the latest Walt Disney. It's a story ABOUT a wolf. Oh that one. Fantasia.

(39) Make a credit card donation. Or for more information ABOUT our work ring …

(40) It's one of the well known things ABOUT money.

However, a great majority of the nouns in these schematic constructions fit into one of the categories listed in Table 6, which suggests that the noun position is not totally schematic in terms of semantic content.

Table 6. Semantic sets of nouns preceding *about* in schematic constructions (type C)

Mode of communication	*information, news, talk, article, conversation, adverts, call, letter, book, drama, film, music, myths, novel, papers, poem, song, story, tale*
Mental state or activity	*dream, nightmares, beliefs, conclusion, confidence, confusion, curiosity, dreams, excitement, fantasy, feelings, idea, ideas, misconstruction, mistake, misunderstanding, pleasure, thinking, thought, trauma, worries, worry*
Opinion or communicating opinion	*argument, comments, discussion, complaint, opinion, advice, apology*

Most of the nouns in the mainly substantive constructions (type A) also fit into a small number of similar categories, exemplified in Table 7.

Table 7. Semantic sets of nouns preceding *about* in substantive constructions (type A)

Mental state or activity	*ideas, think, clue, minds, mind*
Opinion or communicating opinion	*row, qualms, pity, shame*
General nouns*	*fuss, trouble, ways, part, thing, point*

* General nouns are "a small set of nouns having generalized reference within the major noun classes". They are less specific in their reference than other nouns, but they are not as grammatical as the pronouns. (Halliday & Hasan, 1976: 275). General nouns are salience reducing features of the referential kind (Hudson, 1998: 111).

There are some interesting points here:

1. In the category 'opinion or communicating opinion' all the nouns in the dataset have the same negative polarity that we found in the ADJ+*about* dataset.
2. In the substantive group we find a frequent occurrence of general nouns, which are a typical feature of more or less fixed expressions (Hudson 1998: 111–14).
3. The categories overlap to a great extent with those found in the VERB+*about* dataset (see below).

5.3 VERB+*about*

In the same sub-corpus (BNC s-conv) there are 5786 occurrences of VERB+*about*, with 513 different verbs. We have, so far, found some indications of a strong correlation between the sense of *about* and the patterns around the sequence.

As Table 8 shows, only 105 occurrences could not be described in more or less constructional terms.

Table 8. Patterns around VERB+*about* in the BNC s-conv data

Sense of *about*	Tokens	Patterning
approximately	380	Patterns to the right with nominal expressions of time or quantity
around	409	All in phrasal verbs denoting aimless or undesirable motion or action
concerning	3300	All following verbs of communication or mental/emotional state (ratio ~ 50 : 50)
(other 1)	1592	All following *be, have, do, go, get*
(other 2)	105	No discernable patterning, or unanalyzable
Total	**5786**	

The VERB+*about* dataset has not yet been investigated in detail, but it is notable that the semantic categories preceding *about* in the sense of 'concerning' are also those found in the NOUN+*about* data, as mentioned above. In the sense of 'around', *about* only occurs in phrasal verbs, which are substantive constructions. In all of the occurrences in our data, we found the same negative orientation ('aimless or undesirable motion') as in the ADJ+*about* set.

At this point we ask ourselves, could it be that there is something inherent in the meaning of *about* that constrains its potential to form affinities with items preceding it? We think so.

6. Concluding remarks

The general framework of Construction Grammar proves to be a useful tool in the investigation of formulaic language around the relater *about*. It provides both a concrete methodology and a theoretical base that are coherent with our understanding of formulaicity in language. Other signs of formulaicity, and even incipient fixedness, are evident in the abundance of salience reducing features in many of the constructions that have emerged in the course of the study so far, such as: polysemy, decategoriality, non-salient reference, subjectification, and (in the case of very frequent types) reduction.

We have found that something in the region of 80% of the occurrences of *about* ('concerning') in the ADJ+*about* and NOUN+*about* datasets can be described in terms of constructions – from the more substantive and highly idiomatic (*thing about* X *is*, *sorry about that*) to the more schematic ([N] *about*) where the noun

belongs to one of a few sets (general noun, noun of mental state or activity, noun of opinion or communicating opinion). In the case of general nouns preceding *about*, we are not surprised to find that these turn up in the more substantive constructions, given that general nouns provide less specific reference and are thus salience reducing features, which typically occur in more (rather than less) formulaic language (section 2, above).

There are indications of both similarities and differences between the datasets that we have looked at so far:

1. Mental states and activities, opinions, and communication seem to be prevalent semantic sets that immediately precede *about* ('concerning') both as nouns and verbs.
2. The ADJ+*about* set, and the N+*about* set with nouns denoting opinion or communication of opinion, pattern strongly with meanings with a negative or generally unfavourable orientation.
3. We also find this negative orientation in the phrasal verbs, where the sense of *about* is 'around'.

Turning briefly to the learner perspective, mentioned in the opening lines of this paper: There is an abundance of usage guides for the items that we group together under the label 'relater' (e.g., CCEG1;[12] Lindstromberg 1997; Chalker 1999; Carter & McCarthy, 2006). A good deal of what we have discovered in this study is there, between the lines, both in the lists of words that are provided as frequent collocates of *about* and in attempts to systematically describe the semantics of *about*. But these accounts are fragmented, none of them offering a unified description of usage. Several standard works today mention, for example, the use of *on* as an alternative to *about*, such as:

> *About* seems still to be marginally more suitable in formal discourse though it is my impression that *on* is being used more and more often with this meaning […] the cause may be *about's* loss (in the minds of many native speakers) of the literal meaning 'around' [and] the loss of potential for *about* to evoke the image of a surrounded Landmark or, more particularly, of a topic "covered from all angles". This leaves *about* with only the bland, image free meaning, 'concerning'. (Lindstromberg 1997: 139)

> Books, articles and discussions can be *about* or *on* something. But *on* suggests a more serious study of the topic. (Chalker 1990: 4)

12. *Collins Cobuild English Guides. 1 Prepositions.*

On this particular issue, we argue that Chalker is closer to the mark, though we further suggest that it might be the pervasive negative prosody of *about* in many constructions that is causing academics to avoid the word.

But our results so far have more important implications. Current descriptions treat the relaters according to their role or function in traditionally defined constituents – most often as a preposition in prepositional phrases, or as an adverb or particle in phrasal and prepositional verbs. We have found substantial patternings beyond these constituents. Our closing hypothesis is that, since relaters by definition relate two entities, there will be stronger or weaker affinities towards both, each with its own particular constraints. In the case of *about, a book (for example) can potentially be about anything under the sun, hence the Y element in [X about* Y] is far less constrained than the X element, as our results so far have shown. The analysis model that we suggest in this paper might facilitate a more comprehensive description of the relaters and contribute, ultimately, to a better understanding of the formulaic nature of English.

References

Brinton, Laurel. 1996. *Pragmatic markers in English.* Berlin: Mouton de Gruyter.

Brugman, Claudia. 1983 [1988]. Story of *over.* MA thesis, University of California, Berkely. (Published as *The story of over: Polysemy, semantics, and the structure of the lexicon.* New York NY: Garland)

Bybee, Joan L. 1998. The emergent lexicon. In CLS 34: *The panels,* M.C. Gruber, D. Higgins, K.S. Olson & T. Wysocki (Eds), 421–435. Chicago IL: Chicago Linguistics Society.

Carter, Ronald & Michael McCarthy. 2006. *Cambridge Grammar of English*. Cambridge: CUP.

Collins Cobuild English Guides. 1: *Prepositions.* John Sinclair (Ed.), London: Collins.

Chalker, Sylvia. 1990. *English grammar: Word by word.* Surrey: Thomas Nelson & Sons.

Croft, William & D. Alan Cruse. 2004. *Cognitive linguistics.* Cambridge: CUP.

Fillmore, Charles J., Paul Kay & Mary C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language* 64: 501–538.

Goldberg, Adele. 1995. *Constructions.* Chicago IL: University of Chicago Press.

Halliday, Michael A.K. & Ruqaiya Hasan. 1976. *Cohesion in English.* London: Longman.

Hopper, Paul J. 1991. On some principles of grammaticalization. In *Approaches to grammaticalization,* Vol I, E.C. Traugott & B. Heine (Eds), 17–35. Amsterdam: John Benjamins.

Hopper, Paul. 1998. Emergent Grammar. In *The new psychology of language,* Michael Tomasello (Ed.), 155–175. Mahwah NJ: Lawrence Erlbaum Associates.

Hudson, Jean. 1998. *Perspectives on fixedness: Applied and theoretical*. Lund: Lund Universtiy Press.

Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalisations: The What's X doing Y? construction. *Language.* 75(1): 1–33.

Lehmann, Christian. 1995 [1982]. Thoughts on grammaticalization. München: Lincom. (Lehmann (1995) is essentially a reproduction of Lehmann's unpublished (1982) writings)

Lindstromberg, Seth. 1997. *English prepositions explained*. Amsterdam: John Benjamins.
Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Language and communication, J.C. Richards & R.W. Schmidt (Eds), 191–227. London: Longman.
Sinclair, John. M. & Ronald Carter. 2004. *Trust the text: Language, corpus and discourse.* London: Routledge.
Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge MA: Harvard University Press.
Traugott, Elizabeth C. 1982. From propositional to textual and expressive meanings: some semantic-pragmatic aspects of grammaticalization. In *Perspectives on historical linguistics,* W.P. Lehmann & Y. Malkiel (Eds), 245–271. Amsterdam: John Benjamins.
Traugott, Elizabeth C. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper given at the Twelfth International Conference on Historical Linguistics, Manchester, UK, August 1995. (Version of December 1995.)
Tyler, Andrea & Vyvyan, Evans. 2003. *The semantics of English prepositions. Spatial scenes, embodied meaning and cognition.* Cambridge: CUP.
Wiktorsson, Maria. 2003. *Learning idiomaticity: A corpus-based study of native-like idiomaticity in advanced EFL writing*. Stockholm: Almqvist & Wiksell International.

The aim is to analyze NP

The function of prefabricated chunks in academic texts

Elma Kerz & Florian Haas
RWTH Aachen University/Free University of Berlin

Abstract

In the present study we investigate the use and function of prefabricated chunks in academic writing by focusing on what we will term "research predicates", i.e., high-frequency lexical items designating the research process with its key stages. We conducted a manual analysis of these predicates in the academic subcomponent of the British National Corpus and extracted a set of partially lexically filled constructions. Adopting a usage-based constructionist approach and examining its ability to study prefabricated chunks in the register of academic writing, we show that research predicates are part of more complex partially substantive constructions which commonly occur in the register of academic texts and have a acquired a more or less formulaic status. The function of these constructions is to mirror the key phases of an idealized research process.

1. Introduction

The growing interest in the study of what have been commonly referred to as "formulaic sequences", i.e., various kinds of strings of linguistic items which seem to be holistically stored and retrieved from memory, results from an increasing awareness of the pervasiveness of ready-made (prefabricated) expressions within a

language system and the crucial role they play in first and second language acquisition, as well as in language production. Various studies (cf. Schmitt 2004; Wray 2000; Wray & Perkins 2000) have shown that knowledge of formulaic, i.e., preconstructed expressions, plays a vital role for a language user. An array of preconstructed sequences can aid fluent communication since this part of the language is already preassembled and can be easily processed.

Academic writing is conventionalized to a high degree. As Manning and Schütze (1999: 9) point out, a convention is "simply a way in which people frequently express or do something, even though other ways are in principle possible". A high degree of conventionalization of academic texts is particularly evident in its formulaic and systematic makeup, displaying fewer possibilities in the way the research process is presented to the reader. One of the results of the relatively high standardization of the academic way of representing facts (especially in the natural sciences and engineering disciplines) is the reduction of the number of linguistic constructions used for realizations of contents typical of academic texts. This makes academic texts particularly amenable to analysis in terms of frequently recurring ways, i.e., entrenched patterns.

In the present study we will investigate the use and function of prefabricated chunks by focusing on what we will term "research predicates", i.e., high-frequency verbs and the corresponding deverbal nominalized forms designating the research process with its key stages (e.g., *study, research, survey, investigate, investigation, analys/ze, analysis* etc.). We will discuss an array of partially lexically filled constructions involving one of the research predicates in academic writing.

2. A constructionist approach to formulaic sequences

In the relevant literature on prefabricated chunks we encounter a wide range of terms used to refer to these chunks. Wray (2000: 465) provides a list of around fifty terms for describing various facets of formulaicity. One of the reasons for the proliferation of these terms is the diversity of language-related disciplines such as corpus linguistics, descriptive linguistics, lexicography, second language research, description of special-purpose language and foreign-learner language. A further reason lies in the fact that within descriptive linguistics a variety of approaches can be identified, which in turn are associated with different theoretical traditions. These various theoretical traditions have tended to focus on different types of prefabricated chunks. Furthermore, there has been no consensus concerning what criteria are essential for identifying and characterizing

formulaic sequences. Certain criteria have figured more or less within specific theoretical and methodological approaches to "formulaicity", a corollary of which is that these approaches have consequently placed their focus on certain types of prefabricated chunks and simultaneously downgraded other types to a marginal role. In this paper we argue that with the emergence of construction grammar there is a possibility to account for different types of prefabricated chunks within a single theoretical framework.

Construction grammar originally emerged from the necessity to explain idiomatic expressions such as *to come out of the closet on sth,* or *to riot away one's day*. The analysis of idioms led to the rethinking of the syntactic representation that was laid down in the generative framework since the semantic and syntactic unpredictability of idiomatic constructions posed a problem for the generative theoretical framework. A very general notion of "construction" encompasses various types of patterns which may be distinguished in terms of schematicity or abstractness and syntagmatic complexity ranging from simple words to complex constructions and simultaneously displaying various degrees of schematicity, ranging from fully lexically filled and partially lexically filled constructions to fully abstract constructions.

A construction is taken to be a schematic one if it involves grammatical categories such as NP or subject, whereas it is considered substantive if given slots of a schematic construction are filled by specific lexical items. For instance, an expression such as *in the final analysis* is a fully lexically filled, or substantive construction, in which each element is a concrete lexical unit, whereas abstract structural configurations, such as the caused-motion construction [Subject [Verb Object Directional]], are highly schematic constructions.

The scope of a construction can be a single word, where the conventionalization concerns a single frame and its canonical expression (e.g. the verb *analyse*), or it can represent a more complex pattern which must integrate various components (e.g., the so-called *way*-construction, as in *Bill belched his way out of the restaurant*).

Some linguists argue that the essential criterion of "constructionhood" (cf. Goldberg 1995) is the semantic unpredictability of the whole from its component parts, while many others (cf. Langacker 1987, 1991; Croft 2001; Tomasello 2003; Bybee 1985, 1995, 2001; Barlow & Kemmer 2000) argue that patterns are stored and retrieved as wholes even if they are fully predictable, as long as they occur with sufficient frequency (usage-based models). The latter criterion of sufficient frequency is well known under Langacker's notion of "entrenchment". When a complex structure comes to be manipulated as a prefabricated congregation no longer requiring conscious attention to its parts of their arrangement, it acquires

the status of a construction, i.e., it becomes entrenched in the mind of a language user. Langacker uses the "scaffolding metaphor" to describe this process:

> […] component structures are seen as scaffolding erected for the construction of a complex expression; once the complex structure is in place (established as a unit), the scaffolding is no longer essential and is eventually discarded (Langacker 1987: 461).

The effects of these mechanisms can be seen in the effortless, online constructions of complex chunks of language, in accordance with well-established rules. The notion of entrenchment is closely related to those of "automatization" and "habituation". All of these concepts refer to a general psychological mechanism that does not only impinge on the use of language, but also on many other activities in domains such as music or sports.

The notion of entrenchment plays an important role in the phenomenon of "formulaicity". Entrenchment of different kinds of constructions leads to their holistic storage and retrieval, which is commonly considered as one of the decisive criteria for the identification of formulaic sequences. Wray's (2000) definition of the formulaic sequences, for instance, goes as follows:

> [A] sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray 2000: 465)

Holistic storage and retrieval of semantically transparent and syntactically regular constructions in the mind of a language user proves to be perfectly compatible with one of the major tenets of construction grammar, viz. that language system is not entirely free of redundancy and not maximally economic since it simultaneously incorporates both schematic structural configurations and prefabricated chunks of concrete expressions that occur with sufficient frequency in everyday language situations. According to construction grammar, these types of expressions are accessed immediately and easily by language users without an accompanying activation of the corresponding higher-level schematic constructions.

The identification of patterns that exhibit some kind of semantic opacity or syntactic irregularity seems to be a straightforward task. It is clear that these patterns cannot be generated by rules, and hence need to be stored and retrieved as whole. However, as aptly observed by Wray (2000), this type of pattern constitutes a rather small group. As she goes on to say, in order to cover the whole range:

> […] it is necessary to allow for the possibility that word sequences may be formulaic even though they do not need to be, that is, even though they are semantically transparent and syntactically regular. (466)

The same should hold for the identification of constructions within the framework of construction grammar. We think that both criteria, i.e., unpredictability and sufficient frequency can figure when identifying a construction, but that at any one time only one can be decisive for a structural configuration to be regarded as a construction. We are also aware of the fact that it is difficult to precisely define what is meant under the criterion "sufficient frequency". How high has the number of occurrences of a specific pattern to be in order to be regarded as a construction? However, even if there is no "rigorous operationalization of a sufficient frequency threshold" (Gries 2008), it is self-evident that frequency has an impact on the language user's repository of constructions.

Bybee (2006) provides empirical evidence for the assumption that specific instances (exemplars or instantiations) of constructions are part of the cognitive representation of language, and that frequency of use has an impact on the nature of the repository of constructions. She presents various degrees of frequency effect depending on the extent of the frequency (Bybee 2006: 537):

i. low levels of repetition lead to conventionalization only (as in prefabs and idioms)
ii. higher levels of repetition can lead to the establishment of a new construction with its own categories
iii. extremely high frequency leads to the grammaticization of the new construction, the creation of grammatical morphemes, and changes in constituency

The first observation is of crucial importance for constructions of the type presented in this study. We will argue that particular instances of partially lexically filled constructions are registered in linguistic memory indexed with their implications and contexts of use, and that this is especially true for the language users being particularly active within a specific language domain such as that of academic writing.

Within construction grammar the focus has hitherto been on either rather abstract or schematic constructions, such as the resultative or the caused-motion construction, or on fully lexically filled constructions such as idioms or frozen collocations. The role of partially lexically filled constructions involving one or more flexible slot has been reduced to the study of partially substantive idioms (e.g., *jog ⟨someone's⟩ memory, under the auspices of* NP). The constructionhood of idioms is characterized in terms of the aforementioned criterion of semantic unpredictability. However, as we will show in this study, there is a range of partially lexically filled constructions which are semantically predictable involving one or more flexible slots. These slots are constrained as to their lexical filling, i.e., they prefer certain lexical items. The use of research predicates in the register of academic writing

is a case in point. These predicates commonly enter a set of partially lexically filled constructions whose slots are filled by a set of lexically homogenous units that can be described in terms of Fillmorean frame elements (http://framenet.icsi.berkeley.edu/). These elements should be seen as register-specific. They can be regarded as generalizations of the lexical fillers in the partially lexically filled constructions presented in this study. For instance, "research nouns" tend to enter complex noun phrases that can be represented in the following way:

(I) [(NP1's) (AP) N2 [$_{PP}$ *of* (*the*) N3]]NP

Represented in terms of register-specific frame elements, this schema displays the following internal structure:

(II) ⟨Scholar⟩$_{NP\text{-}POSS}$ ⟨Evaluation⟩$_{AdjP}$ ⟨Method⟩$_{AdjP}$ research noun *of* ⟨Object Scope[1]⟩$_{NP}$

The schema contains optional slots signaled by parentheses in (I). In this way, instantiations of that schema do not necessarily involve all of the elements, but combinations of some of them. Some instances of this semi-schematic construction are:

(1) … a conservative, behaviouristic analysis of poverty. (BNC: HP2: 272)
(2) … a psychological investigation of his methods of information processing, … (BNC: CM2: 128)
(3) … sophisticated computational and statistical analysis of data (BNC: A0T: 1223)
(4) Young and Willmott's detailed empirical study … (BNC: F9S: 1282)
(5) the most recent study of Neratius' regulae (BNC: B2P: 47)

They commonly include the *s*-genitive which indicates the scholar conducting a study, a prenominal evaluative adjective (e.g., *conservative, sophisticated* and *detailed*), a "temporal adjective", i.e., an adjective expressing temporal information (e.g., *recent*), an adjective denoting an academic discipline (e.g., *psychological*), an adjective denoting a method/methodology used (e.g., *behaviouristic, computational and statistical, empirical)*, and the postmodifying *of*-phrase identifying the Object Scope, or a combination of these elements. In some cases, the research noun is followed by a *by*-phrase indicating the scholar (e.g., *A large prospective study by Wald et al.* […] (BNC HU3 4957); *A fairly dispassionate analysis by Best (1980)* […] (BNC CHC 91)).

3. Methodological issues

As noted earlier, some approaches define prefabricated chunks entirely in terms of statistical frequency measurement and hence avoid the problem of identify-

1. We will use the label "Object Scope" to refer to an entity or phenomenon under investigation.

ing criteria for the classification and description of these units. They focus on the extraction of word combinations in a corpus on the basis of their frequency and probability of co-occurrence. They regard various theoretical attempts to classify phraseological units as unreliable in contrast to the statistically verifiable output of a purely corpus-driven approach. Even though it is important to conduct statistical analyses of corpora, one of the disadvantages of the exclusively statistical method is that a range of unusual constructions may be overlooked. Before performing a rigorous statistical analysis of corpus data, it is essential to conduct a preliminary analysis based on a manual random-sample survey of concordance lines. This leads to the first rough insights into linguistic phenomena. Instead, we propose to study the meaning potential and the use of lexical units based on a manual analysis of a smaller set of randomly selected corpus data. Subsequent studies using inductive statistical methods may provide further insights.

As this study will demonstrate, there is a range of partially lexically filled constructions in academic texts, i.e., constructions simultaneously including lexically fixed and flexible slots, which cannot be extracted by a simple frequency-driven method. The randomly selected concordance lines were manually inspected in order to filter out the constructions within which the research predicates occurred.

Any study that adopts a usage-based constructionist approach to language, and hence assumes a syntax-lexicon continuum, must take into account partially substantive constructions. Many terms used to refer to prefabricated chunks (cf. Wray 2000) such as "lexical bundles" (Biber 2006) subsume only one type of construction, viz. fully lexically fixed or substantive constructions, which occupy one of the end points of the lexicon-syntax continuum. But, as the study of research predicates clearly indicates, in academic texts we encounter a range of templates consisting of both categories as well as lexical units. Constructions of the type discussed here can be conceived of as patterns involving variables (or "placeholders") that can be filled by certain types of elements.

As a consequence of what was mentioned earlier, we argue for a register-specific notion of entrenchment: some constructions are frequent in a certain domain of linguistic activity only, and they become entrenched for the users of linguistic domains, i.e., registers. A register or genre imposes restrictions on the way information may be organized, and even on what kind of information may be conveyed more easily. Over time, among the scientific community a general consensus arises as to which are the best of currently available formulations.

There have been various corpus-based register studies (cf. Biber, Conrad et al. 1994; Kittredge & Lehrberger 1982; Hunston & Sinclair 2000), which clearly indicate that there is no such thing as "English as a whole" and that

> any patterns generalized for all of English are not likely to be valid for any actual text or register – rather, generalized patterns would merely level the important patterns of use found across registers. Furthermore, we have illustrated the way

> in which these registers/patterns can be interpreted functionally, in terms of differing communicative goals and characteristics of each register (Biber et al. 1999: 82).

The inventory of formulaic expressions undergoes constant change, i.e., it is a dynamic system constantly changing to meet the needs of the language user (Wray 2002: 101). Since a single language user tends to be active within certain domains of life, the arsenal of formulaic expressions should meet the needs of the language users within these domains. As observed by Taylor (2002):

> There will obviously be differences between speakers with respect to which chunks have been committed to memory. Nevertheless, within a given speech community, there will be significant overlap with respect to what has been memorized. (545)

We think that this overlap is particularly large between language users that are active within one domain, such as the academic one.

For the purposes of our investigation we used the academic subcomponent of the British National Corpus. It consists of approximately 15.5 million words and covers a range of academic disciplines, as well as various text types such as research articles, dissertations or textbooks. According to David Lee's genre classification scheme (http://homepage.mac.com/bncweb/manual/genres.html), the academic subcorpus of the BNC consists of texts originating from six different academic disciplines (see Table 1).

Table 1. Structure of the extracted academic subcorpus of the BNC based on Lee's categorization scheme

Domain: academic prose	**No. of words**	**Percentage of the sub-corpus [%]**
humanities and arts	3,321,867	21,53%
medicine	1,421,933	9,22%
natural sciences	1,111,840	7,21%
politics, law, education	4,640,346	30,07%
social & behavioral sciences	4,247,592	27,53%
technology, computing, engineering	686,004	4,45%
Sum	15,429,582	100,00

Partially lexically filled constructions involving research predicates as presented here are common in academic texts of the six disciplines listed in Table 1.

Due to the unavailability of software packages that would allow automatic extraction, the identification of partially lexically filled constructions containing research predicates was largely carried out manually. We inspected sets of ran-

domly selected concordance lines of five research predicates, viz. *study, investigate, research, analys/ze,* and *explore,* displayed in a KWIC (Key Word in Context) format revealing the type of construction in which these lexical items tend to occur.[2]

4. Case study: The use of research predicates in English academic texts

4.1 The register of academic writing

Academic texts are amenable to an analysis in terms of frequently recurring patterns for several reasons. Firstly, their high degree of conventionality has certain effects on the repository of linguistic constructions available to the writer. We have certain expectations as to the way information is presented in academic texts. They include an accurate use of register-specific formulaic sequences.

In contrast to the register of conversation, which is characterized by its online production, the authors of academic texts can take their time when selecting among the options provided for encapsulating information. Academic texts are not produced spontaneously, i.e., constructions are not assembled on the spot but "are carefully planned, edited, and revised" (Biber et al. 1999: 23). This also affects the inventory of linguistic constructions available to the writer. An interesting observation made by the authors of *Longman Grammar of Spoken and Written English* is that despite the fact that native language users "are less consciously aware of register distinctions, it turns out that grammatical differences across registers are more extensive than those across dialects" (Biber et al. 1999: 21). As they go on to say, when language users "switch between registers, they are doing different things with language, using language for different purposes, and producing language under different circumstances" (Biber et al. 1999: 21).

One might argue that whereas in spoken registers one of the major functions of prefabricated chunks is the reduction of online processing, in academic writing one of the main functions is to contribute to the overall impression of a formal,

2. Besides SARA, we also made use of the VIEW interface to the BNC (http://corpus.byu.edu/bnc/). It enabled us to search for patterns involving specific lexical items such as the research nouns *analysis* or *study* in combination with more abstract syntactic categories like verbs. However, one of the major disadvantages of available concordance software packages including the VIEW interface is that they are not able to extract patterns which involve more than one optional slot. For instance, in the construction [NP1 (of NP2) V *that*-clause] the second NP is optional. This phrase may be filled by a very complex noun whose extraction would then necessitate more than five optional slots: *the efficiency of systems of communication …* This phrase may be filled by a very complex noun whose extraction would then necessitate more than five optional slots.

impersonal style. Academic English is characterized by particular academic conventions such as the avoidance of personal language, judgmental words or emotive language. Academic texts are thus produced under highly controlled and edited circumstances, in which the author deliberately signals the impersonal, technical and formal style of this register. The need for a precise and dense packaging of information, as well as for planning and editing, is one of the distinguishing features of this text type. Awareness of these features triggers the use of prefabricated chunks. Partington aptly notes that "in very many genres of writing, pre-cooked expressions are still diagnostic, vital elements" (1998: 20). The author of a certain text deliberately uses such preconstructed patterns in order to signal the register.

Furthermore, academic texts – regardless of which text type (e.g., article monograph) – take a completed or ongoing research process with its key phases as their major topic, i.e., the pragmatic background of their production, namely a completed or ongoing research process and the academic discussion about it, constitutes an essential part of their content (cf. Meyer & Kerz 2004). This increases the probability of encountering formulaic sequences within the register. As Kuiper observed (2004):

> formulaic performance is only possible in routine contexts, i.e., in situations where there is an expectation that things will happen in much the same way that they have happened before. (39)

Biber et al. (1999) observe that the lexico-grammatical patterns:

> found in newspaper articles are quite different from those found in conversation, because, as already shown, speakers in conversation typically have quite different communicative purposes from the writers of newspapers reports. (13–14)

It is rewarding for the study of a specific genre to determine the degree of entrenchment of a construction in question within a specific domain, since the probability of encountering highly entrenched linguistic patterns is raised by frequently and regularly recurring routines typical of specific domains. Domain-specifically entrenched constructions emerging from such a study, due to their lower frequency in the language at large, would otherwise be missed.

Various works on the research process (cf. Bouma & Ling 2004; Weidenborner & Caruso 2005; Gray & Malins 2004) posit distinct phases to describe what in Lakoff's (1987) terms would be called an "idealized cognitive model" of the research process. By comparing these studies and identifying commonalities between them, we were able to deduce a model of the research process with six key phases. According to this model, the research process starts with the definition of a problem and ends with the communication of findings. It is important to bear in mind that

the different stages of the research process do not generally follow each other in a linear sequence, but "are rather part of a continuous iterative cycle, or helix, of experience (consistent with Kolb's 1984 'experiential learning cycle')" (Gray & Malins 2004: 12).

An important characteristic of academic texts is the explicit indication of the aforementioned stages by certain sets of constructions in different sections of the text. This has been analyzed in terms of "generic moves" (Swales 1990), "schematic structure" (Martin 1989), or "generic structural potential" (Hasan 1989), reflecting the conventionalized structuring of genre determined by its communicative purpose.

In what follows we will first outline an idealized cognitive model of the research process with its key phases. As a next step we will present common partially substantive constructions around research predicates that relate to these research phases. As metioned above, an idealized model of the research process, taking into account frequently recurring processes and associated participants, resembles the Lakovian idealized cognitive model (ICM). The latter can be conceived of as an organized abstract framework of objects and relations. Although in reality the research process does not proceed from one stage to another, academic texts give the impression of consisting of the phases temporally following one another. The model hence schematizes what goes on in the research process.[3]

The academic sub-component of the BNC does not exclusively comprise research articles, but also textbooks, dissertations, etc. Despite this diversity of text types, it is possible to deduce the following key phases of the research process usually addressed in academic texts:

- defining the scope and objectives of the study (phase 1);
- constructing or developing a theoretical framework (phase 2);
- employing a convenient method for obtaining explicit solutions/results
- (phase 3);
- finding results (phase 4);
- drawing conclusions (phase 5);
- communicating the findings (phase 6).

---

3. In the EAP/ESP literature we find extensive studies on the disciplinary rhetoric of academic writing which use Swales' "move analysis" (2004), identifying so-called "generic moves". The focus has been on the genre of research articles, which – because of their rather inflexible organization (as required by journals) – prove to be particularly amenable to the analysis in terms of generic moves.

4.2 Formulaic sequences containing research predicates

In what follows we concentrate on the use of prefabricated chunks around research predicates such as *analyse, study, examine,* or *investigate.* The reason for selecting this type of lexical items is that they have the meaning potential to designate the entire research cycle with its key phases. They are high-frequency lexical items in the register of academic writing and belong to the group of so-called "specialised nontechnical lexis" (Cohen, Glasman, Rosenbaum-Cohen, Ferrara and Fine 1988).[4]

A number of corpus-based studies (cf. Biber et al. 1999; Oakey 2002; Biber et al. 2004) of academic prose have highlighted the pervasiveness of an EAP-specific phraseology characterized by multiword combinations that are semantically compositional and syntactically regular (e.g., *as a result of, it is likely that, it has been suggested, as shown in fig, the aim of this study*). These strings are built around lexical items and serve the rhetorical functions prominent in academic writing, viz. signaling the relevant parts of the text.

However, as shown by the examples above, EAP ("English for Academic Purposes") and ESP ("English for Specific Purposes") research on formulaic sequences has hitherto put the emphasis on sequences of three or more fixed lexical items. Constructions dealt with in this paper are syntagmatically complex entities which include slots with specific lexical items, optional slots, slots with abstract categories, as well as slots for a homogenous set of lexical items. Some of the slots are highly abstract or schematic (e.g., NP, *that*-clause), while others are more or less severely constrained with respect to their lexical filling.

Let us now present some partially lexically filled constructions containing research predicates. Combining the automatic queries conducted with the help of Davies VIEW interface to the BNC with the manual inspection of randomly selected concordance lines, we have extracted the following formulaic sequences:

(III) [det *aim/objective/purpose* (*of* NP1)] [*be to* $V_{\text{research verb}}$ NP2],

4. In the EAP/ESP literature on academic vocabulary a distinction between "general service" or basic vocabulary, technical vocabulary and sub-technical vocabulary, is commonly made. The latter, to which research predicates belong, are variously referred to as "frame words" (Higgins 1966), "academic vocabulary" (Martin 1976; Coxhead 2000), for instance. They include lexical items that occur more frequently in academic texts than in non-academic texts and do so consistently across different disciplines and discourse registers without being domain-specific. The investigation of this type of vocabulary does not require the specialist knowledge of the relevant technical domain's conceptual content. In other words, sub-technical vocabulary is not specific enough in meaning to belong to the terminology of a specific discipline, but is simultaneously more formal than "general" English.

whereby the N slot of NP1 is usually filled by the research noun *study* or *analysis* and NP2 denotes a phenomenon under investigation, (III) is a construction commonly used in the first phase of the research process, viz. identifying and formulating a viable research question, as in (6) and (7):

(6) The purpose of this study is to investigate the ability of adsorbents, administered enterally, to reduce systemic endotoxaemia in a hapten induced model of colitis. (BNC: HU2: 773)

(7) The aim is to analyse a problem which economic growth alone has failed to cure – […] (BNC: AS6: 4)

A further construction commonly used to signal the first phase of the research process is the one in (IV), exemplified in (8):

(IV) [$NP1_{\text{research noun}}$ *be designed to* $V_{\text{research verb}}$ NP2],
whereby the N-slot within NP1 is usually filled by the research noun *study,*
whereas the N-slot of NP2 is commonly filled by one of the research verbs *investigate* or *examine.*

(8) This study was designed to investigate the possibility that transferring to human insulin has a direct effect on the perception and experience of hypoglycaemia.

The query "[nn*] [vb*] *designed* to" in the VIEW interface yielded that *study* is the most frequent noun (n=21) filling the N-slot, followed by *procedure.* (n=6).

When the researcher informs the reader about the breadth or scope of his or her research endeavor, they usually make use of the following type of construction:

(V) $NP1_{\text{research noun}}$ V $NP2_{\text{Object Scope}}$

The V slot is here filled with the following verbs: *include*, *involve*, *center around*, *focus*, as well as their negative counterparts *neglect*, *exclude*, etc.

One may argue that an alternative standard way of specifying and narrowing down the scope of the research endeavor is through the use of complex PNP ("P" standing for a preposition) constructions such as *in terms of*, *with regard to*, *with respect to* or *in the light of*, or the preposition *for* in combination with one of the research predicates:

(VI) $NP1_{\text{Object Scope}}$ *be examined for/in terms of/with regard/respect to*/etc. $NP2_{\text{Parameter}}$

Here "NP1" denotes an entity under investigation, and "NP2" names the condition under which this entity is examined, as in (9)–(11):

(9) Sections were examined by light microscopy for bacterial, protozoal, fungal and viral enteric pathogens. (BNC: HU2: 1575)

(10) … discrimination was examined in terms of the extent to which performance on each item differentiated between the normal and language-disordered children. (BNC: CG6: 176)

(11) … , articulation is examined with respect to three word positions and with respect to production within sentences. (BNC: CG6: 1126)

As shown by the examples above, the frame element "parameter" in academic texts is commonly introduced by the following expressions: *for, with regard/respect to, in relation to, in terms of.* When entering a complex transitive construction of the type in (9), research verbs denote an activity of accessing the presence or the absence of certain phenomena in the entity examined.

Analyzing data by using a specific method constitutes the third phase of the research process. Relevant constructions are (VII) and (VIII):

(VII) $\text{NP1}_{\text{Scholar}}$ *use/perform* $\text{NP2}_{\text{Method}}$ *to* $\text{V}_{\text{research verb}}$ $\text{NP3}_{\text{Object Scope}}$

(VIII) $\text{NP1}_{\text{Method}}$ *be used/applied/performed to* $\text{V}_{\text{research verb}}$ $\text{NP3}_{\text{Object Scope}}$ (the passive variant).

(12) Whitfield (1979) has used this correlational approach to analyse the functions of the auditory cortex. (BNC: CMH: 605)

(13) Stathmokinetic and immunohistochemical techniques were used to study the effect of 1,25 (OH) 2D3 and its analogues on cell proliferation in human rectal mucosa and a colon cancer cell line. (BNC: HU4: 1989)

The query "[v*] *to analy?e*" yielded that the V-slot is most frequently filled by the verb *used* (13 occurrences). Crucially, we obtain different results if we conduct the same query for the other registers of the BNC: for the spoken component, for instance, the query yields "have to analyse" (2), whereas for fiction it is "trying to analyse" (8). If we take a closer look at the concordance lines in which the sequence "used to analyse" occurs, we notice that it is part of the larger partially substantive construction in (VIII), where NP1 is filled by a restricted set of nouns denoting a method or procedure (e.g., *stathmokinetic and immunohistochemical techniques*) that an implied scholar uses in order to examine an entity or phenomenon. In the "Results" sections, research nouns are the most salient lexical units, functioning as the subject of the verb which expresses reporting results in the type of construction in (IX):

(IX) $\text{NP1}_{\text{research noun}}$ [V NP2/*that*-clause/*wh*-clause]VP

Some instances of this construction are given in (14)–(16):

(14) The study revealed variations in attitude and usage pattern as between the various social classes. (BNC: G3F: 1294)

(15) Moreover, the present study indicated that the syn-PLA2 and cat-PLA2 values of patients with a necrotising form of acute pancreatitis had a tendency to remain increased for a longer time than the values in patients with oedematous acute pancreatitis. (BNC: HWS: 7054)

(16) A study of Cheshire in the first quarter of the fifteenth century, however, shows how the social ties of the gentry class were rooted in the locality. (BNC: HWG: 1180)

The most common verbs filling the slot of the construction in (IX) are: *indicate, show, reveal, suggest*. *Reveal*, for instance, occurs 158 times in the academic sub-component of the BNC. In 72 instances, its subject is one of the research nouns, as in (17)–(18):

(17) Further research revealed that what had at first appeared to be a bizarre anomaly was in fact a cultural feature shared by many different Indian people. (BNC: CS0: 364)

(18) Restriction analysis of the rescued plasmid revealed that it had the expected structure. (BNC: K5Y: 336)

The manual analysis of randomly selected sentences including research nouns showed that when filling the NP1 slot in the construction [NP1 V *that*-clause] they commonly co-occur with result verbs, i.e., verbs such as *show* or *reveal.* The query "analysis [v*] that" in VIEW, for instance, yielded the following results: the most frequent verb filling the V-slot is the verb *show* (*showed* 21 + *shows* 14 times), followed by the verbs *suggest, indicate* and *reveal.* In the case of the noun *study,* the most common verb is also *show* (*shows* 49 + *showed* 48 + *show* 25), followed by *suggest(s)* (57), *indicate* (29) and *find* (18). The noun *research* is most frequently followed by *suggest,* followed by *show*, *indicate* and *find* in terms of frequency. The noun *survey* is found with the following verbs: *show*, *find*, *suggest*, *report*, *reveal* and *indicate*. Hence, the verbs filling the V-slot of the construction in (IX) make up a semantically rather homogeneous group of lexical items.

The fifth stage of the research process involves "the critical synthesis of the whole experience, demonstrating its value and significance through effective communication and dissemination" (Gray and Malins 2004: 15). We will call the move addressing this stage "underlining the effects of study".[5] In this stage the researcher draws conclusions about his research findings and provides their value and significance to the wider research context. The constructions in (X) and (XI) are commonly used:

(X) $NP1_{\text{research noun}}$ [V NP2]

5. The move "underlining the effects of study" is not necessarily aligned to one of the key phases of the research process.

(XI) NP1 [V-ed (AP) *from* $NP2_{research\ noun}$]

Typical verbs occupying the V-slot of the patterns found in the context "underlining the effects of study" are: *raise, contribute, lead to, enable NP to, emerge from, obtain from, derive from, glean from,* etc. Most of these verbs are found in the list constituting what Hunston and Francis (2000) refer to as the logical relation "be result of".

(19) The studies led to the revelation of reasons why the programme was less successful … (BNC: CED: 236)

(20) … research has helped to increase the number of women subjects in mainstream European and North American psychology, and the range of topics over which they are studied. (BNC: CMR: 512)

As the examples above illustrate, "research nouns", viz. the nominalized forms of research verbs, are often found in combination with verbs of facilitation or causation (e.g., *allow, enable,* or *provide*), which "indicate that some person or inanimate entity brings about a new state of affairs" (Biber et al. 1999: 363).

We are aware of the fact that the division between "showing results" and "effects of study" is not always easy to draw. Nevertheless, the latter is used to refer to situations where the scholar goes beyond presenting results and points out effects these results had or might have on further research.

5. Conclusion

As shown above, the term "construction" has the potential to cover a wide range of formulaic sequences including highly schematic, abstract structural configurations, semi-schematic ones, partially filled as well as prefabricated chunks of concrete expressions that occur with sufficient frequency.

The existence of partially lexically filled constructions of the type discussed here supports one of the main assumptions made by construction grammar, viz. that of the syntax-lexicon continuum. We argued that this type of construction is particularly well-established in specific registers or genres. Although their extraction from corpus data is time-consuming, the study of such constructions is of great importance, as they constitute an integral part of the inventory of constructions language users have at their disposal. It is hoped that progress in computational linguistics will finally help us automatically distill constructions of the type presented in this paper.

The research process incorporates several key stages or phases and academic texts usually address these various phases. As a consequence of the relatively high

standardization of the scientific (or "academic") style of representing facts (this specifically holds for natural science and engineering disciplines), we observe a small number of typical patterns for communicating the various stages of the research process. In other words, a rather limited set of constructions within academic texts signal certain communicative-semantic moments within the research process. As we have shown, research predicates commonly enter a set of partially lexically filled constructions, which in turn are part of so-called moves reflecting the key phases of an idealized research process. Apart from their text-structuring function, we pointed out that the relevant constructions are used by authors of academic texts to signal the register itself.

Although we have pointed out the utility of usage-based constructionist approaches to the study of prefabricated chunks, we are also aware that it is important to study formulaic sequences from different theoretical perspectives and to apply different kinds of methodological techniques. In general one should hope that in the future there will be more cross-fertilization between different linguistics disciplines as well as various theoretical frameworks and methodologies and that these disciplines can profitably benefit from each other. The UWM Linguistic Symposium on formulaic sequences was a case in point.

References

Barlow, Michael & Suzanne Kemmer. 2000. *Usage-based models of language*. Stanford CA: CSLI.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2004. If you look at … : Lexical bundles in university teaching and textbooks. Applied Linguistics 25: 371–405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Bouma, Gary D. & Rod Ling 2004. *The research process*, 5th Edn. Melbourne: OUP.

Bybee, Joan L. 1985. *Morphology*. Amsterdam: John Benjamins.

Bybee, Joan L. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10: 425–455.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: CUP.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 529–551.

Cohen, Andrew, Hilary Glasman, Phyllis R. Rosenbaum-Cohen, Jonathan Ferrara & Jonathan Fine. 1988. Reading English for specialised purposes: Discourse analysis and the use of student informants. In *Interactive approaches to second language reading*, P.L. Carrell, J. Devine & D.E. Eskey (Eds), 152–167. Cambridge: CUP.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213–238.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: OUP.

Gray, Carole & Julian Malins. 2004. *Visualizing research: A guide to the research process in art and design*. Ashgate: Aldershot.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago IL: University of Chicago Press.

Gries, Stefan T. 2008. Phraseology and linguistic theory: A brief survey. In *Phraseology: An interdisciplinary perspective,* S. Granger & F. Meunier (Eds), Amsterdam: John Benjamins.

Hasan, Ruqaiya. 1989. The structure of a text. In *Language, context, and text: Aspects of language in a social-semiotic perspective* (68), M.A.K. Halliday & R. Hasan (Eds), Oxford: OUP.

Higgins, John J. 1966. Hard facts. *ELT Journal* 21(1): 55–60.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hunston, Susan & John Sinclair. 2000. A local grammar of evaluation. In *Evaluation in text: Authorial stance and the construction of discourse*, S. Hunston & G. Thompson (Eds), 74–101. Oxford: OUP.

Kittredge, Richard & John Lehrberger (Eds), 1982. *Sublanguage: Studies of language in restricted semantic domain*. Berlin: de Gruyter.

Kolb, David A. 1984. *Experiential learning: Experience as the source of learning and development*. Upper Saddle River NJ: Prentice-Hall.

Kuiper, Koenraad. 2004. Formulaic performance in conventionalized varieties of speech. In *Formulaic sequences: Acquisition, processing and use*, Norbert Schmitt (Ed.), 37–54. Amsterdam: John Benjamins.

Lakoff, George. 1987. *Women, fire, and dangerous things*. Chicago IL: University of Chicago Press.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar*, Vol. 1: *Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 1991. *Foundations of cognitive grammar*, Vol. 2.: *Descriptive application*. Stanford CA: Stanford University Press.

Manning, Christopher & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge MA: The MIT Press.

Martin, Anne V. 1976. Teaching academic vocabulary to foreign graduate students. *TESOL Quarterly* 10(1) : 91–97.

Martin, J.R. 1989. *Factual writing: Exploring and challenging social reality*. Oxford: OUP.

Meyer, Paul Georg & Elma Kerz. 2004. Towards a conception of lexical pragmatics. In *Anglistentag 2003 München: Proceedings*, C. Bohde, S. Domsch & H. Sauer (Eds), 97–111. Trier: Wissenschaftlicher Verlag.

Oakey, David. 2002. Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In *Using corpora to explore linguistic variation*, R. Reppen, S.M. Fitzmaurice & D. Biber (Eds), 111–129. Amsterdam: Longman.

Partington, A. 1998. *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Schmitt, Norbert (Ed.), 2004. *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Swales, John M. 1990. *Genre analysis: English in academic and research settings*. Cambridge: CUP.

Swales, John M. 2004. *Research genres: Exploration and applications*. Cambridge: CUP.

Taylor, John. 2002. *Cognitive grammar.* Oxford: OUP.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge MA: Harvard University Press.

Weidenborner, Stephen & Domenick Caruso. 2005. Writing research papers: A guide to the process. New York NY: St. Martin's Press.

Wray, Alison & Mick Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20(1): 1–28.

Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21(4): 463–489.

Fixedness in Japanese adjectives in conversation

Toward a new understanding of a lexical ('part-of-speech') category*

Tsuyoshi Ono & Sandra A. Thompson
University of Alberta/UC Santa Barbara

*We thank Joan Bybee, Patricia Clancy, Robert Englebretson, Yumiko Konishi, Junko Mori, Masayoshi Shibatani, and especially Shoichi Iwasaki and Yasuhiro Shirai for their valuable comments on the ideas in this paper; we are responsible for the way in which we've chosen to incorporate their input. We are grateful to the audience members in the Rice University Linguistics Department for their helpful input. We would also like to thank the participants of the 'little experiment' in our study for their willingness to help and John Fry for making some quantitative figures available to us.

Abstract

Japanese adjectives have received a fair amount of attention for their intriguing morphological and diachronic properties. Adjectives have also been discussed in the typological literature, largely in terms of their status as a lexical category vis-à-vis nouns and verbs. Rather little research has been done, however, on the everyday use of adjectives in Japanese conversation. In our paper, we aim to show that (a) adjective usage in conversation is intricately bound up with fixedness and frequency; (b) a usage-based approach reveals that interactional and cognitive practices are deeply intertwined in this lexical category for Japanese speakers; (c) these facts reflect the nature of human language as an emergent phenomenon. Based on a substantial corpus of Japanese conversations, we find that (a) attributive adjectives are very rare; (b) among predicative adjectives, as well as the rare attributive adjectives, the most frequently occurring forms strongly tend to be associated with various types of fixedness, demonstrating its central status in our attempt to represent the grammar for real speakers.

1. Introduction

Our paper is a case study of a lexical category: adjectives in Japanese. Specifically, we examine the type of adjectives which have traditionally been called *i*-adjectives. Our investigation probes the 'psychological reality' of the category 'adjective' for Japanese speakers, and what this category tells us about 'formulaic language'. We will argue in favor of these two hypotheses:

Hypothesis 1. Conversational Japanese strongly favors PREDICATE adjectives over ATTRIBUTIVE adjectives.

Hypothesis 2. Whether predicative or attributive, an understanding of Japanese adjectives in everyday talk involves various facets of FIXEDNESS.

Hypothesis 2a. ATTRIBUTIVE and PREDICATIVE adjectives in Japanese show DIFFERENT TYPES OF FIXEDNESS.

Hypothesis 2b. ONGOING LEXICALIZATION is a prominent feature of Japanese adjective usage

2. Theoretical background

We take a strongly empirical, usage-based, approach to grammatical structure (Barlow and Kemmer 2002; Tomasello 2003; Bybee 1998; 2006, 2007). We further align with an emerging research tradition within a usage-based framework, which is coming to be known as 'interactional linguistics' (Ford 1993; Ford et al. 2002a,b, Ford & Thompson 1996; Fox 1987, 1995, 2001; Fox et al. 1996; Selting &

Couper-Kuhlen 2001; Hakulinen & Selting 2005; Ochs et al. 1996; Thompson & Couper-Kuhlen 2005). We engage with the issue of the 'open choice' vs. 'idiom' principles (Erman & Warren 2000; Pawley & Syder 1983; Sinclair 1991; Wray 2002), asking, with these researchers, "what are the respective roles of 'open choice' vs. 'learned as a chunk' in spoken language use?" Here we will argue that 'learned as a chunk' plays a much larger role in the use of adjectives in Japanese than has been assumed in the literature, whether traditional, pedagogical, functional, or generative. In fact, our findings suggest that 'fixedness' plays a much larger role in the representation of a lexical category, in this case the 'adjective' category, than we've seen discussed in the literature for any language (with the notable exception of Hopper 1997a,b, discussing the emergence of the category 'verb').

Linguistic categories (what grammarians once called 'parts of speech') have been argued for on the basis of structural properties inferred from constructed examples (e.g., Schachter 1985). While this approach has proven to be highly useful in relying on structural, rather than semantic, grounds for determining lexical category boundaries, it has the drawback of not being based on speakers' actual behavior. Our paper is one attempt to re-examine the category of 'adjective' in Japanese in the context of everyday conversational interactions. Our investigation has shown that the actual nature of 'adjective' in Japanese interaction is rather different from what has been widely assumed.

3. Previous research on Japanese adjectives

The category of adjective is arguably THE major issue in lexical categories in Japanese grammar; every sizeable reference grammar or pedagogical grammar of Japanese discusses it.

3.1 Traditional approaches

Within a structural tradition, research on Japanese has been almost exclusively concerned with the categoriality of 'adjective' in Japanese; that is, is there such a category, and if so, what forms should be considered as belonging to this category? Research representatives of the traditional approach written in English includes Backhouse (1984, 2004), Kuno (1973), Iwasaki (2002), Jorden & Noda (1987), Martin (1975), and Shibatani (1990).

Within this tradition, two types of 'adjectives' have been recognized:

(a) *i*-adjectives (*keiyooshi*)
= 'inflected adjectives'

(b) *na*-adjectives (*keiyoo-dooshi*)
= 'adjectival nouns'/'nominal adjectives'[1]

It is essentially uncontroversial that there are substantial differences in general between the two subclasses. Table 1, based on Backhouse (2004), summarizes the literature:

Table 1. Differences between *i*-adjectives and *na*-adjectives

	***i*-adjectives**	***na*-adjectives**
morphology	resemble **verbs** : they inflect	resemble **nouns**: they don't inflect[2]
syntax	can be directly used attributively	can only be attributive with postposition *na*
	cannot be used with copula *da*	can be used with copula *da*
diachrony	its antecedent found in the earliest records	generally accepted as coming from nouns

Backhouse (2004: 51) concludes that:

(a) Both subclasses of adjective can be clearly distinguished from verbs and nouns respectively.
(b) There are strong arguments for treating them both as types of adjectives in Japanese.

It is noteworthy that in this literature, we easily find constructed examples of both attributive and predicative adjectives, e.g., (Shibatani 1990: 216):

predicative:

(1) ano hito wa **kirei** da[3]
that person TOP pretty COP
'that person is pretty'

---

1. As the terms given in (6) suggest, the status of the *na*-adjectives has long been contested: are they **nouns** or are they **adjectives?**

2. Intriguingly, the lack of inflection appears to be the motivation for adjectives borrowed from another language to be automatically placed into the *na*-adjective class (Backhouse 2004; Iwasaki 2002):

awesome na
yucky na
cool na (found on the internet)

3. Please see the appendix for transcription and glossing conventions.

attributive:

(2) **kirei** na hito
pretty person
'a pretty person'

However, these researchers don't indicate any possible usage difference between them.

In the generative tradition, 'predicative' and 'attributive' have been considered as two 'open slots', defining the category of 'adjective'. Positing these slots could then support the generative metaphor, whereby a concise set of structural rules with lexical items filling the 'slots' would produce an infinite number of 'novel sentences'.

These traditional structuralist claims may be valuable for providing hypotheses about the cognitive representation of the adjective category, but we question whether they reflect what speakers know about 'adjectives' as they engage in their everyday interactions.

3.2 Usage-based approaches to Japanese adjectives

Two notable studies of Japanese adjectives, which can be seen as precursors to the more recent usage-based approaches to grammar, have been carried out by Uehara (1996, 1998). In contrast to the traditional consensus discussed just above, Uehara (1996) argues that the Noun vs. Nominal Adjective distinction is so fuzzy that *na*-adjectives should be considered a non-prototypical subclass of NOUN. We will come back to this shortly.

Uehara (1998), following Croft 1991, 2001, suggests that in fact lexical categories be determined on the basis of the constructions in which they occur (cf. also Stefanowitsch and Gries 2003 for the same point). We fully agree with this suggestion, and will show that this is highly relevant to fixedness.

As far as we know, there has not been any research written in either English or Japanese looking at Japanese adjectives based on actual interactional data.

4. Data

The data for this study consist of 16 naturally occurring conversations among friends, family members, and couples, totaling about 100 minutes and consisting of more than 7000 Intonation Units (Du Bois et al. 1993). The conversations were among between 2 and 5 speakers of Standard Japanese. The transcripts have been produced and checked by teams of native speaker transcribers.

5. Methodological procedures and overview

5.1 Form rather than function

The database for this study was constructed by extracting both *i*-adjectives and *na*-adjectives from transcripts while listening to the recording. Specifically, we identified relevant examples and determined their function and their fixedness by inspecting specific interactional contexts in transcripts while repeatedly listening to the original recording.[4] This yielded a total of about 600 *i*-adjectives and relatively clear instances of *na*-adjectives combined.[5] We'll call this our "*i/na* database".

5.2 ***na***-adjectives and ***i***-adjectives

Our data confirm Uehara's 1996 finding that it's generally difficult to distinguish *na*-adjectives and nouns. For example, the following stems are found in our data:

(3) *tanoshimi* 'fun'
okanemochi 'rich'
mania 'obsessive, fanatical'

These adjectives can be followed by *na*, in which case they would be categorized as *na*-adjectives. But they can also be followed by the genitive particle *no*, in which case they'd be nouns. On the internet, they are found in both forms:

(4) – both *tanoshimi na* (*na*-adj) and *tanoshimi no* (noun)
– both *okanemochi na* (*na-adj*) and *okanemochi no* (noun)
– both *mania na* (*na-adj*) and *mania no* (noun)

As noted, such pairs support Uehara's claim; they also support the prevailing hypothesis that *na*-adjectives have developed, and are still developing, from nouns.[6] Because of this indeterminacy resulting from a grammaticization process in progress, we have chosen to focus on *i*-adjectives, for which there is general agreement as to their categorial status as adjectives. As we will see in 6.2.2 below,

---

4. Based on our years of experience transcribing and using transcripts, we advocate this admittedly time-consuming and labor-intensive approach to study linguistic structure. Transcripts, even after multiple checkings, can only represent a fraction of what goes on in actual interaction; the recording gives us a much better representation of what humans do. In our view, watching or listening to the original recording is the very first step in an attempt to understand the interaction between form and function in the use of spoken language.

5. These include 426 instances of *i*-adjectives; the rest were *na*-adjectives.

6. Shoichi Iwasaki (p.c.) reminds us of the suggestion that even the now more established adjective class of *i*-adjectives may have a nominal origin (see Okamura 1968).

however, there is also indeterminacy associated with the class of *i*-adjectives, since new members constantly keep being added to that category through lexicalization. Out of the 600 or so adjectives, our "*i*-adjective" database was constructed, consisting of **426** *i*-adjectives.[7] From here on, then, when we say 'adjective', we mean '*i*-adjective'.

5.3 Frequency

According to Fry (2003), the six most frequent predicate *i*-adjectives in Japanese 'CALLHOME' data were those shown in (5):[8]

(5) *ii* 'good' (by far the most frequent, as in our data)
warui 'bad'
takai 'tall, big'
ookii 'big'
yasui 'cheap'
ooi 'many'

Comparing these with the most frequent adjectives in our database of 426 *i*-adjectives, we find the results given in (6):

(6)	*ii* 'good'	169
	warui 'bad'	20
	ikenai 'bad'	19
	chiisai 'small'	12
	wakaranai/wakannai 'ununderstandable'	12
	sugoi 'great'	11
	kawaii 'cute'	10

Clearly, *ii* 'good' is also our most frequent adjective by far, with 169/426; consequently *ii* will be featured below in our discussion of the role of frequency and fixedness.

6. Our findings

Recall that we are focusing for the remainder of our paper on just the *i*-adjectives; we now turn to the empirical support for our claims.

---

7. These figures do not include adjectives used adverbially such as *sugoi* 'awfully'. They also do not include forms marked with the productive derivational suffixes *-tai* 'want', *-ppoi* 'look like', etc., which conjugate like i-adjectives.

8. The high frequency of *yasui* 'cheap' and *takai* 'tall, big' (which we note also means 'expensive') is almost certainly due to the nature of the CALLHOME data, where speakers talk about the no-cost phone calls the corpus builders let them make.

6.1 Claim 1: Predicate adjectives outnumber attributive adjectives

Table 2 shows that conversational Japanese strongly favors predicate adjectives over attributive adjectives.

Table 2. Predicative vs. Attributive Adjectives

Predicative	373	**88%**
Attributive	53	12%
Total	426	100%

6.2 Claim 2: Whether predicative or attributive, an understanding of Japanese adjectives in everyday talk involves various facets of fixedness

Even impressionistically, Japanese adjective use seems to involve much more fixedness than in English. To be sure, a degree of attributive adjective fixedness has been suggested for English by Chafe (1982), (1994: 118–119), Chafe and Danielewicz (1987), and Englebretson (1997). But such a claim has not been made for English predicate adjectives as far as we know. We will provide examples and return to a comparison with English below.

In this paper, we recognize that 'fixedness' is a multidimensional property, and that fixedness can be measured, even with a relatively small corpus such as ours (Erman & Warren 2000; Wulff 2007).

In this section, we will demonstrate that Japanese adjective use seems to involve **much more fixedness than in English**. We will see that the data show a heretofore unrecognized **variety of types of fixedness** in the use of Japanese adjectives.

As expected, there's a strong correlation between fixedness and **frequency** in adjective use in Japanese (see Bybee 1998, 2001a,b, 2002a,b, 2006, 2007 for frequency effects on fixedness).

6.2.1 *Claim 2.1: Attributive and predicative adjectives in Japanese show different types of fixedness.*

A. Attributive adjectives

Recall that attributive adjectives make up only 12% of our conversational data. Both English and Mandarin conversations show higher percentages of attributive adjective use, suggesting the Japanese may be unusual in this regard.[9] Among these 12%, a prevailing pattern obtains.

---

9. Two studies for English suggest attributive adjectives make up about 36% of all adjectives; see section 6.2.1.A below. For Mandarin, preliminary results suggest the percentage of attributives is about 29%; see Thompson and Zhan (forthcoming).

Light heads

Where attributive adjectives do occur, the majority (75% (40/53)) are associated with some type of fixedness and most of them (38/40) in a specific way: they occur with a semantically light head, or 'generic noun', such as those shown in (7).[10]

(7) – *koto, mono, no* 'thing, stuff, matter'
– *uchi* 'period'
– *toki, koro* 'time'
– *hoo* 'direction'
– *toko(ro)* 'place'
– *yatsu* 'guy'
– *hito* 'person'

Examples of the light head NPs preceded by attributive adjectives include these from our database:

(8) **oishii** **mono** atta shi sa
delicious thing exited and FP
'(they) had delicious things.'

(9) **hidoi** **yatsu** mo iru mon da ne
awful guy Mo exist NOM COP FP
'awful guys exist/there are awful guys!'

(10) **chitchai** **koro**
little time
'a little time/when (I was) little'

(11) **isogashii** **toki**
busy time
'a busy time/when (you are) busy'

(12) **yowai** **no** ooi n da kedo
weak one many NOM COP but
'weak ones are many/(there) are many weak ones.'

(13) **tooi** **toko** itta kara[11]
distant place went so
'(I) went to a distant place (far away school) so.'

(14) **ii** **hito** ga ireba na
good person GA exist.if FP
'(it would be good) if there is someone good (= 'partner' or 'lover').'

10. Some of these nouns have been discussed in Japanese linguistics as *keishikimeishi* ('formal nouns'); cf., e.g., Masuoka and Takubo (1992).

11. *toko* is a reduced form of *tokoro.*

(15) **umai** **koto** yaru
skillful thing do
'(there might be people who) are successful at it)' (very idiom-like)

These examples illustrate the pervasive tendency for the relatively infrequent attributive adjectives in our conversational corpus to occur in a fixed expression with a light head. Most of these examples feel somewhat lexicalized. In particular (14) is clearly lexicalized,[12] and (15) is very idiom-like. Furthermore, there is evidence that these light heads tend to occur only with a modifier, such as an attributive adjective or a relative clause. This evidence takes two forms. First, intuitively, it is difficult to imagine many of these light heads on their own except in certain fixed expressions. Second, though our database is relatively small, it contained only a few instances of these light heads occurring without a modifier.[13] It is intriguing to note that Ozeki and Shirai (2005, 2007), studying children's acquisition of what have been taken to be 'relative clauses', found that:

> The prototypical relative clauses that Japanese children use early on modify generic nouns and pronouns such as *mono* 'thing', *tokoro* 'place', *-no* 'one' with stative/generic predicates … , as in this example:
>
> (9) (Sumi 2;03)
> *[Kore ireru] mon doko ni aru?*
> This put.in thing where Loc exist
> 'Where is the thing [in which I put this]?' (2007: 247)

Ozeki & Shirai argue that

> Japanese children's relative clauses are extended from adjectival modification, which they have already acquired. (2007: 247)[14]

These facts suggest that speakers draw on a set of 'light head' construction templates for constructing attributive adjective + NP phrases. In this set are included a general template with a broad type frequency, something like:

---

12. Some dictionaries list *ii hito* as a separate entry, indicating its fully lexicalized status.

13. We found two instances of independent *mono*, one instance of independent *tokoro*, and several instances of independent *hito*, though independent *hito* included some semi-fixed/written like expressions.

14. Ozeki & Shirai (2007) take their results to be strong support for the position persuasively articulated by Matsumoto (1988, 1997) and Comrie (1998a, b) that Japanese noun-modifying clauses form a continuum with **adjectival modification**, and are thus not structurally appropriately described in terms of 'relative clauses', but as 'modifying clauses simply attached to the head noun' (p. 248).

(16) [ADJ + N_{light}]

and a set of specific templates with a more restricted type frequency for the ADJ slot:

(17) ADJ + *koto*
ADJ + *mono*
ADJ + *no*
ADJ + *hito*
etc.

The fixedness profile for the relatively few attributive adjectives in the database, then, involves a constructional template with a light noun as its head.

Comparison with English

Turning now to attributive adjective usage in English, we are aware of three studies that have investigated adjective usage in English conversations: Chafe (1982), Thompson (1988), and Englebretson (1997). Comparing the figures in the Chafe and Thompson papers, Englebretson argues persuasively that genre is a primary determinant of the ratio between attributive and predicative adjectives in everyday talk:

> … interactions where participants are evaluating and commenting on shared referents tend to be heavy on predicate adjectives, while interactions such as narrative or conference [discourse], which involve the introduction of new referents … , tend to be heavier on attributive adjectives. (p. 418).

Since our Japanese database consists of highly participatory interactions with many shared referents, we decided to take as our comparison standard the Thompson (1988) database, consisting of 308 adjectives drawn from audio-recorded conversations among friends, and the four most equivalent conversations from Englebretson's study.[15] These figures are given in Table 3.

Table 3. Percentages of Attributive Adjectives in two studies of English conversation compared to the current study of Japanese conversation

English		**English**		**Japanese**	
(Thompson 1988)		(Englebretson 1997)			
Predicate	Attributive	Predicate	Attributive	Predicate	Attributive
68% (209)	32% (99)	62% (495)	38% (303)	**88%** (373)	**12%** (53)

15. These four conversations were from the Santa Barbara Corpus of Spoken American English Vol. 1 (Du Bois et al. 2000): 'Appease the Monster', 'Actual Blacksmithing', 'Runway Heading', and 'Hey Cutie Pie'. These conversations are audio-taped interactions among friends and family members; the first three are face-to-face, and the fourth is a phone call.

There are two noteworthy facts revealed by Table 3:

a. Table 3 shows clearly that, although predicative adjectives outnumber attributive adjectives in both English and Japanese (and Mandarin, the only other language for which we have quantitative findings, for that matter), the ratio for attributive adjectives is considerably lower for Japanese (12%) than for English (36%).
b. Having shown, then, that, as in Japanese, in everyday English conversations the TOKEN FREQUENCY of attributive adjectives is lower than that of predicative adjectives, we note that the TYPE FREQUENCY of [ADJ + N] NPs in English is relatively high. That is, compared to Japanese, where the range of types of head nouns is severely constrained to light heads, as we have just shown, there is a much larger range of types of head nouns occurring with attributive adjectives in English, as illustrated in these examples, which were picked at random from the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000):

(18) KEN: … Yeah.
I think it'll be real interesting
I think it'll be a **real**,
(H) a **good slide show**. SBC 15: 6.325

(19) MICHAEL: **creative people** generally do what they love to do.[16] SBC 17: 29.395

(20) LENORE: he's having **bad luck** with that car. SBC 06: 59.88

(21) MONTOYA: if you're the chairperson of u=m .. a **major corporation**?
SBC 12: 72.53

(22) SHARON: because their [parents,]
CAROLYN: [**bi=g mistake**].
SHARON: were too lazy to come,
… and,
.. and fill out the **stupid form**, SBC 04: 206.22

In fact, [ADJ + N] NPs which could be argued to be compositionally understood, of the type so readily found in English conversations, such as *a good slide show* or *big mistake*, are markedly rare in our Japanese data; we found only 13 clear examples, that is, 25% of the 53 attributive adjectives and 3% of the total 426 *i*-adjectives.

16. One might consider *people* in (19) as a light head in English. Its Japanese equivalent *hitobito* does not feel similarly light. This is probably because, unlike English *people*, *hitobito* is a reduplicated form, not a base form of the word. In fact, our informal experiment below seems to support the 'heaviness' of *hitobito*.

Intriguingly, however, compositionally understood [ADJ + N] NPs appear with surprising regularity in Japanese language textbooks and linguistics articles based on constructed examples, as illustrated by these examples:

(23) *akarui heya* 'bright (sunny) room'
ii beddo 'good bed'
ookii oshiire 'big closet'
atarashii terebi 'new TV'
furui isu 'old chair' (Makino et al. 1998: 68–70)

We hypothesize that the discrepancy between what is found in these sources and what our data reveal is due to a combination of both a 'written language bias' (Linell 2005) and influence from English and/or English-based linguistic theories in linguistics and language pedagogy.

Such illustrations of apparently freely productive [ADJ + N] combinations are typical in Japanese language pedagogy and sometimes result in a curious situation: Ono has noticed that Japanese language students often produce [ADJ + N] NPs which are rather non-native-like. For instance, the following example is from a written piece by an English-speaking second-year student:

(24) *?ooi*[17] *tiishatsu*
many T-shirt

Such an NP strikes native speaker ears as distinctly odd. As reasonable as 'interference errors' like this might be, student-created NPs like this further support the finding from our data that attributive-adjective NPs can't be as freely created in Japanese as they are in English.[18]

As a final empirical confirmation of the difference between the two languages, we conducted a very informal little experiment. Ten native speakers of Japanese were asked to rate six [Adj + N_{light}] NPs from our conversation data and translated versions of six English examples given in (18) – (22) above. The 12 examples were presented in a randomized order, and the speakers were asked to select one of the following:

1 = more natural
2 = OK
3 = less natural

17. Unlike the English quantifier *many*, Japanese *ooi* is a clear-cut *i*-adjective.

18. In fact, if we translate (18) – (22) into Japanese, they sound like 'translationese', a variety of Japanese which is widely known in Japan due to the heavy influence of English, both written and spoken, and other European languages. (e.g., Yanabu 1982 and Morioka 1988).

The rationale behind this experiment was if the celebrated generative rule of modifying nouns with prenominal adjectives in Japanese is anything close to what speakers actually operate with in real life, the Japanese versions of (18)–(22) based on such a rule should be as acceptable as examples taken out of actual use. Table 4 shows the scores for actual examples from our conversation data; we see that the speakers found these light-head NPs quite acceptable with these attributive adjectives.[19]

Table 4. Naturalness scores for 6 NPs from our conversational database

	Examples from our data	**mean score**
oishii mono	'delicious thing'	1.2
hidoi yatsu	'awful guy'	1.0
tooi tokoro	'distant place'	1.0
chiisai koro	'when (I was) little'	1.3
ii hito	'good person'	1.0
isogashii toki	'when (I was) busy'	1.1

Table 5 gives the scores for translated versions of the English examples given in (18)–(22). In Table 5, we can see that these NPs, where the heads are not light, are rated much lower.

Table 5. Naturalness scores for translations of 6 English attributive adjectives with non-light heads

ii suraidoshoo	'good slide show'	1.6
soozooteki na hito	'creative people'	2.0
warui un	'bad luck'	2.6
shuyoo na kaisha	'major corporation'	2.3
ookii machigai	'big mistake'	2.1
baka na mooshikomi yooshi	'stupid form'	2.8

Attributive adjectives: summary

We have seen, then, that attributive adjectives in Japanese are relatively rare, at least compared to English and Mandarin (see Footnote 8). Furthermore, we have seen that, in comparison to English, Japanese attributive adjectives are measurably more fixed, strongly tending to occur in fixed expressions with light heads.

Let's turn now to the fixedness of predicate adjectives.

19. As we have noted, there are very few attributive adjectives in our data. The examples in Table 4 were chosen to represent the most frequent head types.

B. Predicate adjectives in Japanese

Recall that predicate adjectives are the great majority of adjectives (88%) in our conversational database; we present here some examples from our collection:

(25) kimochi **warui**
feeling bad
'(something is) unpleasant/(one feels) sick!'

(26) (talking about guys having to financially treat their girlfriends)
harawanakya **ikenai**
pay.not.if bad
'if (you) don't pay, (it's) bad/you have to pay'

(27) **sugoi** jan
terrific TAG
'terrific, isn't it!'

(28) (after hearing that the interlocutor had a chance to hear Spanish)
ii nee
good FP
'good!'

(29) **omoshirokatta** yo
interesting.past FP
'(it) was fun!'

Clearly, many of these examples illustrate a very common use of predicate adjectives, namely what Clancy et al. (1996) term 'reactive tokens'. To demonstrate the fixedness of predicate adjective, we present here a case study on the most frequent adjective overall, namely *ii* 'good'.

Table 6 shows the distribution of *ii* 'good' between attributive and predicative uses.

Table 6. Distribution of *ii* 'good' across attributive and predicative uses.

Predicative	159	94%
Attributive	10	6%
Total	169	100%

Fixedness and ii as a predicate adjective

Following Bybee, since *ii* is the most frequent adjective by far in our entire database, we might expect it to show a high rate of fixedness. And indeed, following up on our discussion of attributive adjectives just above, we find that, while 75% of attributive adjectives overall occur with light heads, fully 90% (9 out of the 10) cases of attributive *ii* have light heads, as in (14), repeated here.

(14) **ii** **hito** ga ireba na
good person GA exist.if FP
'(it would be good) if there is someone good (or 'partner', 'lover').'

For predicate adjectives, 67% (106 out of 159) instances of predicative *ii* occur in clearly fixed expressions. Now, again, as we might expect from Bybee's work on frequency and fixedness, with a highly frequent adjective like *ii*, we find various TYPES of fixedness. To illustrate that this is indeed the case for *ii*, we group these expressions into the following three types, based on criteria discussed in the research of Biber et al. (1999: 'lexical bundles'), Bolinger (1976), Bybee (1998, 2001a, 2002a, 2006, 2007), Erman & Warren (2000), Fillmore (1989), Fillmore et al. (1988), Kay & Fillmore (1999), Pawley (2007), Pawley & Syder (1983), Sinclair (1991), Wulff (2007), Wray (2002), and Wray/Perkins 1999):

1. 'Idioms'
2. 'Constructions'
3. 'Fixed to context'

We suggest that these types of fixedness, arising out of the usage-based literature primarily oriented to English, are equally appropriate to capture the nature of adjectives as used and categorized by Japanese speakers in their everyday interactions. We have found that these types are neither mutually exclusive nor exhaustive; they are presented here to give a sense of what the data suggest about the cognitive categories of actual speakers.

Here we will consider relatively clear cases in each of these categories briefly in the context of *ii* 'good' as a predicate adjective.

a. 'Idioms' (8 cases)
We take 'idiom' to characterize expressions that are:

– fully lexically specified
– not fully compositional

In general, if they truly reflect speaker categories, we expect idioms to be listed in dictionaries, and indeed, Japanese dictionaries include many lexicalized expressions involving *ii* 'good', as illustrated in (30):

(30) - *kakko ii* 'appearance good' > ' stylish'[20]
[cf. ?*sugata ii* 'appearance good']
– *atama ii* 'head good' > 'smart'
[cf. ?*noo ii* 'brain good']

20. *kakko* is itself a phonologically reduced form of *kakkoo* 'appearance'.

– *kimochi ii* 'feeling (is) good' > 'feel good'
[cf. ?*kokoro ii* 'heart good']
[cf. *kankaku ii* 'feeling (is) good' > 'has a good sense']
– *tenki ii* 'weather (is) good' > 'sunny'
– *ashiba ga ii* 'foothold (is) good' > 'conveniently located'
[*ashiba ga ii* has 'compositional' and 'non-compositional' features:

– it has the 'nominative' marker *ga* which compositional clauses with S + PredAdj 'should' have[21]
- but it is less compositional, since a normal speaker might not know what *ashiba* means]

b. 'Constructions'

We use the term 'construction' to refer to a large group of semi-fixed patterns with some open and some lexically fixed slots, as illustrated in the following five patterns.

(b.1) Adj or V *hoo ga ii* '(it is) better to …' (9 examples)

The expression *hoo ga ii* '(it's) better to …' is an ideal example of a construction with some open slots and some lexically fixed slots. To give a flavor of the fixedness in this expression, (31a) shows a typical occurrence of *hoo ga ii* from our database:

(31a) isogashii **hoo** **ga** **ii**
busy direction GA good
'the busy direction is good/(it's) better to be busy.'

(31b) shows that substituting *warui* 'bad' for *ii* 'good' results in an expression that seems distinctly odd:

(31b) *isogashii **hoo** **ga** **warui**
busy direction GA bad

However, substituting *dame* 'bad' seems acceptable, though it does not occur in our database:

(31c) isogashii **hoo** **ga** **dame**
busy direction GA bad
'the busy direction is bad/(it's) not good to be busy.'

The following example provides another perspective on the fixedness of *hoo ga ii*. (32a), from our database, contrasts with both (32b) and (32c), where neither *warui* 'bad' nor *dame* 'bad' are likely to be heard:

(32a) shinda **hoo** **ga** **ii** ne
die.past direction GA good FP
'(it's better) to die!'

21. Cf. Ono et al. (2000) for discussion of *ga* as a pragmatic, rather than a case-marking, particle.

BUT:

(32b) ?shinda **hoo** **ga** **warui** ne
die.past direction GA bad FP

(32c) ?shinda **hoo** **ga** **dame** ne
die.past direction GA bad FP

(b.2) *V-ba ii* 'if (one does *V*), (it is) good' →'(one) should (do …)' (15 examples)

Akatsuka (1992) and Clancy et al. (1997) draw our attention to a particularly interesting example of an *ii* construction, namely *V-ba ii* as a hortative 'one should', literally 'if (one does *V*), (it is) good' →'(one) should (do …)'. Their research convincingly reveals the role of the very high frequency of - *ba ii* in language directed to children in language socialization; our data show that *-ba ii* is a recurrent construction in adult interaction as well, as illustrated in (33) and (34a):

(33) koko ni kure-**ba** **ii** n desu ka
here at come-if good NOM COP QUES
'(is it) good if I come here/should I come here?'

(34a) kure-**ba** **yokatta**[22] noni
come-if good.past but
'(you) should have come'

However, (34a) shows that substituting the semantically reasonable *warui* 'bad' for *ii* 'good' in this construction results in an utterance which sounds distinctly odd.

(34a) *kure-**ba** **warukatta**[23] noni
come-if bad.past but

(b.3) *N de(mo) ii* (literally) '(even) with N is good' (35 examples)

The most frequent construction with predicative *ii* in our database is *N de(mo) ii* '(even) with N is good' Examples (35) and (36a) illustrate *N de(mo) ii* '(even) with N is good' from our database:

(35) getsuyoobi **demo** **ii**
Monday even good
'even Monday is good'

(36a) gosenen **de** **ii** to omou no
5000.yen DE good QUOT think FP
'(I) think that 5000 yen is good.'

22. *yokatta* is past tense form of *ii*.

23. *waru-* is a form of *warui* 'bad'.

Attempting to modify (36a) with *warui* 'bad' instead of *ii* 'good' yields an unlikely utterance:[24]

(36b) *gosenen **de** **warui** to omou no
5000.yen DE bad QUOT think FP
'(I) think that 5000 yen is bad.'

Similarly, trying to substitute the negative of *ii*, namely *yokunai* 'not good' also strikes speakers as strange:

(36c) ?gosenen **de** **yo**kunai to omou no
5000.yen DE good.not QUOT think FP
'(I) think that 5000 yen is not good'

The following instances of *N demo ii* '(even) with N is good' are even more idiom-like than those we have just considered:[25]

(37) dotchi **demo** **ii**
which even good
'either is good'

(38) itsu **demo** **ii**
when even good
'any time is good'

(39) doo **demo** **ii**
how even good
'anyway/thing is good'

(b.4) *V-te (mo) ii* 'is good to do' (14 examples)

The construction *V-te (mo) ii* 'is good to do' involves a verb in the non-finite *-te* form followed by *ii* 'good', optionally with *-mo*. As with *N de(mo) ii*, which we considered just above, *V-te (mo) ii* 'is good to do', is also semi-fixed. Thus, we have the following two examples from our data

(40) harawanaku**te** **mo** **ii**
pay.not even good
'(it's) good even not to pay/(you) don't have to pay'

(41a) omae kae**tte** **ii** zo
you go.back good FP
'(it's) good (for) you to go home/you can go home'

---

24. Yasuhiro Shirai (p.c.) suggests that the antonym of *ii* in this use would be *dame* rather than *warui*. As he himself notes, however, the use of *dame* would also yield an unlikely utterance.

25. In fact, as Shoichi Iwasaki (p.c.) has suggested to us, there is a more general construction of the form *Question Word demo ii*, of which these examples are specific instantiations.

But, again, if we considered whether speakers might say (41b), with *warui* 'bad' instead of *ii* 'good', we would find that they would not be likely to do so:[26]

(41b) ?omae kaet**te** **warui** zo
you go.back bad FP
'(it's) bad (for) you to go home/you can't go home'

(b.5) *Adj or V kara ii* 'because Adj or V, (it's) good' (6 examples)

The particle *kara* belongs to a set of clause-final particles marking various types of adverbial clauses, functionally similar to the suffix *-ba*.[27] To a suggestion that the speaker should find a girlfriend in Japan, for example, that speaker said:

(42) iru **kara** **ii** ya
exist because good FP
'(I'm) good because (I) have (one)' =
'(I) don't have to because (I) have (one).'

Examples with *kara ii* illustrate constructions just as do the expressions in b.1) to b.4). But this construction also illustrates our third type of fixed expression exhibited by *ii* 'good', namely expressions that are 'fixed to context', which we turn to next.

c. 'fixed to context' (12 cases)

The expressions we're characterizing as 'fixed' in the sense of 'fixed to context' differ from both what we're calling 'idioms' and what we're calling 'constructions' in that these patterns can be analyzed as 'compositional', but they are conventional for use in more specific discourse contexts.

As one type of example of 'fixed to context' expressions, consider the following example, where *ii* is used to check if it is a good time to talk:

(43) (at the beginning of a telephone conversation)
Y: anata ima ya- **ii** no?
you now good FP
'(are) you good now/can you (talk) now?'
N: **ii** no.
good FP
'(I'm) good/yes'

Though we have no further examples of this particular usage in our database, it is our impression that this is a common way to accomplish this particular task.

---

26. Pat Clancy reminds us that the converse of *ii* in conditionals is not *warui*, but *dame/ikenai* as in *kaetcha dame/ikenai* 'can't go home', which we fully agree with. The point of this example is to show that *V-te (mo) ii* (and its negative converse) is semi-fixed, so that trying to produce a converse by simply replacing a lexical item does not yield an acceptable utterance.

27. We note, however, that they are different morphologically: *kara* attaches to finite forms while *ba* attaches to hypothetical forms (non-finite).

Similarly, *ii* is also used to make an offer, which includes this situation at the dinner table:

(44) (at a dinner table, the hostess asks a guest if she has had enough stew)
moo **ii** no shichuu wa
already good FP stew TOP
'(are you) good with the stew?/(have you had) enough stew?

Interestingly, rejecting an offer is also performed by *ii*.[28]

(45) (N and Y are trying to pick a restaurant for a group dinner. N is talking about sending Y a brochure of one candidate restaurant)
N: atashi ga sono shashin no deta no o okuru kara
I GA that photo of come.out.PAST one O send so
'I will send (you) the one with photos so …'
Y: @@@
N: ne
FP
'OK?'
→ Y: i- **ii** wa yo
good FP FP
'(I'm) good (without it)/(I) don't need (it)'

As can be seen, examples of 'fixed to context' *ii* are typically accompanied by one or more final particles, which suggests that there might be a general template associated with this type, *ii* FP 'good FP'.

Summary: Predicate Adjectives

To demonstrate the fixedness of predicate adjectives in Japanese, we have chosen to focus on the overwhelmingly most frequent adjective, *ii* 'good' as a case study. We have seen that *ii* is used in three familiar types of fixed expressions, 1) 'idioms', 2) 'constructions', and 3) 'fixed to context'.

We turn now to another familiar aspect of fixedness: ongoing lexicalization.

6.2.2 *Claim 2.2: Ongoing lexicalization is a prominent feature of Japanese adjective usage*

As noted above, Uehara (1996) has drawn our attention to *na*-adjectives as an open class. On the other hand, our preceding discussion might have given the reader an impression that *i*-adjectives form a closed class whose membership is easy to

28. This use is probably related to *Adj or V kara ii* 'because Adj or V, (it's) good' discussed in b.5). The major difference is that, unlike the latter, it is clearly associated with a refusal action.

define. This idea that *i*-adjectives form a closed class has been suggested by some researchers (e.g., Uehara 1996). However, our data clearly show that the category of *i*-adjective is also an open class category, due to constant addition of new members. As a case in point, we consider *i*-adjectives with the negative suffix *–nai*. It is well-known in Japanese linguistics that forms with *-nai* inflect just as do *i*-adjectives. Our data show that, as we might expect, we can align these *nai*-forms on a continuum to illustrate that negated predicates with *-nai* are becoming re-analyzed as *i*-adjectives. This further underscores the importance taking account of fixedness in any attempt to understand the lexical category of adjective in Japanese. For convenience, we divide the continuum into two groups as follows.

Recognized in dictionaries as i-adjectives

Some of these *nai*-forms are already fully recognized as *i*-adjectives.[29] That is, though we suspect that many speakers may not consciously think of them as adjectives, these *nai*-forms are listed as adjectives in dictionaries, and their non-negative counterparts often don't exist, as shown in (46):

(46)

– *tsumara**nai*** 'boring' (the parts are recognizable but the meaning of *tsumara* isn't clear)
– *tamara**n(ai)*** 'unbearable' (the parts are recognizable but the meaning of *tamara* isn't clear)[30]
- *ike**nai*** 'bad' (the parts are recognizable but the meaning of *ike* isn't clear)
- *monotari**nai*** 'unsatisfying' (compositional: 'thing-fill-NEG', but very 'idiomatic')
– *tondemo**nai*** 'unbelievable' (morphological breakdown not obvious)
- *shooga**nai*** 'no good' (*shoo*, a reduced form of *shiyoo* 'way/method', cannot appear on its own)
– *doo shiyoo mo **nai*** 'no way' (*shiyoo* 'way/method' isn't used anymore)
– *wakara**nai**/wakan**nai***[31] ('don't understand' → 'forget it')

29. As expected, there is variation as to which forms are recognized as *i*-adjectives among different dictionaries. Forms given in this section are listed as *i*-adjectives in at least one of the dictionaries which we consulted. Since dictionaries are generally conservative in including newly lexicalized forms, we feel justified in assuming that there is a very good chance that forms included in a dictionary are already lexicalized in the language.

30. Both *tamaranai* and *tamaran* are found in our data.

31. The majority have the form *wakannai*.

Recognized, but not as i-adjectives

Somewhat less lexicalized are many *-nai* forms which appear to be following the same path to become new *i*-adjectives, as shown in (47):[32]

(47) - *joodan ja* ***nai*** 'no kidding'
- compositionally, it means 'it is not a joke'
- *kankei* ***nai*** 'irrelevant'
- *imi* ***nai*** 'meaningless'
- *muite****nai*** 'unsuited'
- *shinjirare****nai*** 'unbelievable'

While 'part-of-speech' labels are not given, they are found in some dictionaries, from which we can infer that they are also becoming established as independent lexical items.

Thus, while the above *nai*-forms are not found or recognized as *i*-adjectives in dictionaries yet, we find them in our database being used just as full-fledged predicate adjectives are, namely as reactive tokens, as discussed above in 6.2.1.B. From this usage, we can again draw the inference that speakers are beginning to categorize them as *i*-adjectives:

(48) **shinjirarenai** nee
believe.potential.NEG FP
'unbelievable, you know!' (said about a just-uttered fact)

(49) **kankei** **nai** yo
relationship not.exist FP
'(it) doesn't matter'

Summary: nai-forms

Various shades of fixedness are to be expected in arenas where new adjectives are evolving through ongoing lexicalization. We've shown that *nai*-forms are lexicalizing as *i*-adjectives in just this way, resulting in the degrees of fixedness we see in our conversational database. The category of *i*-adjective thus needs to be understood as an open class category, due to constant addition of new members.

6.2.3 *Summary*

In this section we first showed that Japanese predicate adjectives outnumber attributive adjectives to an extent not found in the other languages for which comparable conversational adjective data are available.

32. We did not include any of these in our counts, as our coding was conservative.

We then showed that Japanese attributive adjective use is associated with a particular type of construction [ADJ + N_{light}], namely, attributive adjectives in conversation strongly tend to occur with light heads.

Focusing next on fixedness in predicate adjectives, we took *ii* 'good' as our 'case study', and showed the range of fixed expressions in which *ii* can be found, following this with a discussion of the fixedness continuum that results from the ongoing lexicalization of *nai*-forms into *i*-adjectives.

7. Conclusions and implications

We take our investigation to have strongly supported the two claims with which we opened this paper. Our data show that:

Claim 1. Conversational Japanese strongly favors PREDICATE adjectives over ATTRIBUTIVE adjectives.

Claim 2. Whether predicative or attributive, an understanding of Japanese adjectives in everyday talk involves various types and degrees of FIXEDNESS.

Claim 2a. ATTRIBUTIVE and PREDICATIVE adjectives in Japanese show DIFFERENT TYPES OF FIXEDNESS.

Claim 2b. Ongoing lexicalization is a prominent feature of Japanese adjective usage

One interpretation of our findings is that the conversational data clearly reveal the category of adjective to be an EMERGENT category.[33] This means that it is dynamic, fluid in its membership, and a by-product of a specific community of humans going about their daily business.

Consistent with a view of linguistic categories as emergent in this sense, then, we have not been surprised to find that various types of FIXEDNESS characterize the category of 'adjective' in Japanese. For the class of Japanese *i*-adjectives, well over half of its conversational tokens occur in fixed expressions, underscoring the insistence of the researchers whose work has inspired our study (cited above in 6.2.1.B.) that linguists regard frequency and degrees of fixedness in everyday language use as central in their attempts to understand linguistic structure. That is, our investigation of the lexical category 'adjective'

33. On the emergence of linguistic structure, see Bybee 1998, 2001b, 2002b, 2006, 2007; Bybee and Hopper 2001; Englebretson 2003; Helasvuo 2001a, b, Hopper 1987, 1988, 1990; Hopper & Thompson 1984, 2008; Huang 1999; Lindblom et al. 1984; Weber 1997; inter alia.)

in Japanese everyday talk reveals a fascinating reality of human language: on the one hand, language is dynamically fluid due to constant on-going change. On the other hand, our investigation highlights a neglected part of the reality, whereby language is massively driven toward fixedness by constant creation of fixed expressions.

List of symbols

@ = laughter	NOM = nominalizer
COP = copula	QUES = question
FP = final particle	QUOT = quote
LOC = locative	TAG = tag
NEG = negative	TOP = topic

References

Akatsuka, Noriko. 1992. Japanese modals are conditionals. *In The joy of grammar: A festschrift in honor of James D. McCawley*, D. Brentari, G.N. Larson & L.A. Macleod (Eds), 1–10. Amsterdam: John Benjamins.

Backhouse, Anthony E. 1984. Have all the adjectives gone? *Lingua* 62: 169–186.

Backhouse, Anthony E. 2004. Inflected and uninflected adjectives in Japanese. In *Adjective classes: A cross-linguistic typology,* R.M.W. Dixon & A.Y. Aikhenvald (Eds), Cambridge: CUP.

Barlow, Michael & Suzanne Kemmer (Eds), 2002. *Usage-Based Models of Language*. Stanford CA CSLI.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Edward Finegan, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English.* London: Longman.

Bolinger, Dwight. 1976. Meaning and memory. *Forum Linguisticum* 1: 1–14.

Bybee, Joan. 1998. The emergent lexicon. *CLS* 34: 421–435.

Bybee, Joan. 2001a. Frequency effects on French liaison. In *Frequency and the emergence of linguistic structure*, J.L. Bybee & P.J. Hopper (Eds), 337–359. Amsterdam: John Benjamins.

Bybee, Joan. 2001b. *Phonology and language use.* Cambridge: CUP.

Bybee, Joan. 2002a. Mechanisms of change in grammaticization: The role of repetition. In *Handbook of historical linguistics*, R.J. & B. Joseph (Eds), 602–623. Oxford: Blackwell.

Bybee, Joan. 2002b. Sequentiality as the basis of constituent structure. In *The evolution of language from pre-language*, T. Givon & Bertram Malle (Eds), 109–132. Amsterdam: John Benjamins.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 529–551 .

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: OUP.

Bybee, Joan & Paul Hopper. 2001. *Frequency and the emergence of linguistic structure* [Typological Studies in Language 45]. Amsterdam: John Benjamins.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In *Spoken and written language: Exploring orality and literacy*, D. Tannen (Ed.), 35–54. Norwood NJ: Ablex.

Chafe, Wallace. 1994. *Discourse, consciousness, and time*. Chicago IL: The University of Chicago Press.

Chafe, Wallace & Jane Danielewicz. 1987. Properties of spoken and written language. In *Comprehending oral and written language*, R. Horowitz & S.J. Samuels (Eds), 83–113. New York NY: Academic Press.

Clancy, Patricia M., Sandra A. Thompson, Ryoko Suzuki & Hongyin Tao. 1996. The conversational use of reactive tokens in Japanese, Mandarin, and English. *Journal of Pragmatics* 26(1): 355–387.

Clancy, Patricia M., Noriko Akatsuka & Susan Strauss. 1997. Deontic modality and conditionality in discourse: A cross-linguistic study of adult speech to young children. In *Directions in functional linguistics*, A. Kamio (Ed.), 19–57. Amsterdam: John Benjamins.

Comrie, Bernard. 1998a. Rethinking the typology of relative clauses. *Language Design* 1(1): 59–86.

Comrie, Bernard. 1998b. Attributive clauses in Asian languages: Towards an areal typology. In *Sprache in Raum und Zeit, In memoriam Johannes Bechert*, Band 2, W. Boeder, C. Schroeder, K.H. Wagner & W. Wildgen (Eds), 51–60. Tübingen: Narr.

Croft, William. 1991. *Syntactic categories and grammatical relations*. Chicago IL: The University of Chicago Press.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective.* Oxford: OUP.

Du Bois, John, Stephan Schuetze-Coburn, Danae Paolino & Susanna Cumming. 1993. Outline of discourse transcription. In *Talking data: Transcription and coding methods for language research*, Jane A. Edwards & Martin D. Lampert, eds., 45–89. Hillsdale NJ: Lawrence Erlbaum Associates.

Du Bois, John W., Wallace Chafe, Charles Meyer & Sandra A. Thompson. 2000. *Santa Barbara Corpus of Spoken American English,* Part One. Philadelphia PA: Linguistic Data Consortium.

Englebretson, Robert. 1997. Genre and grammar: Predicative and attributive adjectives in spoken English. *BLS* 23: 411–421.

Englebretson, Robert. 2003. *Searching for structure: The problem of complementation in colloquial Indonesian conversation.* Amsterdam: John Benjamins.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20(1): 29–62.

Fillmore, Charles J. 1989. Grammatical construction theory and the familiar dichotomies. In *Language processing in social context*, R. Dietrich & C.F. Graumann (Eds), 17–38. Amsterdam: Elsevier.

Fillmore, Charles J., Paul Kay & M. Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64: 501–538

Ford, Cecilia E. 1993. *Grammar in interaction: Adverbial clauses in American English conversations.* Cambridge: CUP.

Ford, Cecilia E., Barbara A. Fox & Sandra A. Thompson. 2002a. Social Interaction and grammar. In *The new psychology of language: cognitive and functional approaches to language structure*, Vol. 2, M. Tomasello (Ed.), 119–143. Mahwah NJ: Lawrence Erlbaum Associates.

Ford, Cecilia E., Barbara A. Fox & Sandra A. Thompson. 2002b. Constituency and the grammar of turn increments. In *The language of turn and sequence*, Cecilia Ford, Barbara A. Fox & Sandra A. Thompson (Eds), 14–38. Oxford: OUP.

Ford, Cecilia E. & Sandra A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the projection of turn completion. In *Interaction and grammar*, E. Ochs, E.A. Schegloff & S.A. Thompson (Eds), 135–184. Cambridge: CUP.

Fox, Barbara A. 1987. *Anaphora and the structure of discourse*. Cambridge: CUP.

Fox, Barbara A. 1995. The category 'S' in English conversation. In *Discourse grammar and typology*, W. Abraham, T. Givón & S.A. Thompson (Eds),153–178. Amsterdam: John Benjamins.

Fox, Barbara A. 2001. On the embodied nature of grammar: Embodied being-in-the-world. In *Complex sentences in grammar and discourse*, J. Bybee & M. Noonan (Eds), 79–100. Amsterdam: John Benjamins.

Fox, Barbara A., Makoto Hayashi & Robert Jasperson. 1996. A cross-linguistic study of syntax and repair. In *Interaction and grammar*, E. Ochs, E.A. Schegloff & S.A. Thompson (Eds), 185–237. Cambridge: CUP.

Fry, John. 2003. *Ellipsis and wa-marking in Japanese Conversation*. Oxford: Routledge.

Hakulinen, Auli & Margret Selting (eds.) 2005. *Syntax and lexis in conversation*. Amsterdam: John Benjamins.

Helasvuo, Marja-Liisa. 2001a. *Syntax in the making: The emergence of syntactic units in Finnish conversation.* Amsterdam: Benjamins.

Helasvuo, Marja-Liisa. 2001b. Emerging syntax for interaction: Noun phrases and clauses as a syntactic resource for interaction. In *Studies in interactional linguistics*, M. Selting & E. Couper-Kuhlen (Eds), 25–50. Amsterdam: John Benjamins.

Hopper, Paul. 1987. Emergent grammar. *BLS* 13: 139–157.

Hopper, Paul, J. 1988. Emergent grammar and the A Priori Grammar constraint. In *Linguistics in context: Connecting observation and understanding*, D. Tannen (Ed.), 117–134. Norwood NJ: Ablex.

Hopper, Paul J. 1990. The emergence of the category 'proper name' in discourse. In *Redefining linguistics*, H. Davis & T. Taylor (Eds),149–162. London: Routledge.

Hopper, Paul J. 1997a. Dispersed verbal predicates in vernacular written narrative. In *Directions in functional linguistics*, Akio Kamio (Ed.), 1–18. Amsterdam: John Benjamins.

Hopper, Paul J. 1997b. Discourse and the category 'verb' in English. *Language and Communication* 17(2): 93–102.

Hopper, Paul J. & Sandra A. Thompson. 1984. The discourse basis for lexical categories in Universal Grammar. *Language* 60(4): 703–752.

Hopper, Paul J. & Sandra A. Thompson. 2008. Projectability and clause combining in interaction. In *Crosslinguistic studies of clause combining: The multifunctionality of conjuntions* [Typological Studies in Language 80]. Amsterdam: John Benjamins.

Huang, Shuanfan. 1999. The emergence of a grammatical category definite article in spoken Chinese. *Journal of pragmatics* 31(1): 77–94.

Iwasaki, Shoichi. 2002. *Japanese*. Amsterdam: John Benjamins.

Jorden, Eleanor Harz & Mari Noda. 1987. *Japanese: The spoken language*, part 1. New Haven CT: Yale University Press.

Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *What's X Doing Y?* construction. *Language* 75: 1–33.

Kuno, Susumu. 1973. *The structure of the Japanese language*. Cambridge MA: The MIT Press.

Lindblom, Björn, Peter MacNeilage & Michael Studdert-Kennedy. 1984. Self-organizing processes and the explanation of phonological universals. In Explanations for language universals, Brian Butterworth, Bernard Comrie & Östen Dahl (Eds),181–203. Berlin: Mouton.

Linell, Per. 2005. *The written language bias in linguistics: Its nature, origins, and transformation.* Oxford: Routledge.

Makino, Seiichi, Yukiko A. Hatasa & Kazumi Hatasa. 1998 *Nakama* 1: *Japanese communication, culture, and context*. Boston MA: Houghton Mifflin.

Masuoka, Takashi & Yukinori Takubo. 1992. *Kiso nihongo bunpoo: Kaiteiban* (Basic Japanese grammar: A revised version). Tokyo: Kurosio.

Matsumoto, Yoshiko. 1988. Semantics and pragmatics of noun-modifying constructions in Japanese. *Berkeley Linguistics Society* 14: 166–175.

Matsumoto, Yoshiko. 1997. *Noun-modifying constructions in Japanese: A frame-semantic approach.* Amsterdam: John Benjamins.

Morioka, Kenji. 1988. *Gendaigo Kenkyuu Shiriizu: Buntai to Hyoogen* (Modern Japanese Studies Series 5: Styles and Expressions). Tokyo: Meiji Shoin.

Ochs, Elinor, Emanuel A. Schegloff & Sandra A. Thompson (Eds), 1996. *Interaction and grammar*. Cambridge: CUP.

Ono, Tsuyoshi, Sandra A. Thompson & Ryoko Suzuki. 2000. The pragmatic nature of the so-called subject marker ga in Japanese: evidence from conversation. *Discourse Studies* 2(1): 55–84.

Okamura, Masao. 1968. *Keiyooshi no katsuyoo no seiritsi* (The establishment of the inflection of adjectives). In *Joodai no Kotoba* (The Language of the Joodai Period), K. Mabuchi (ed.,) 230–245. Tokyo: Shibundoo.

Ozeki, Hiromi & Yasuhiro Shirai. 2005. Semantic bias in the acquisition of relative clauses in Japanese. In *Proceedings of the 29th Annual Boston University Conference on Language Development*, Vol. 2, A. Brugos, M.R. Clark-Cotton & S. Ha (eds.) 459–470. Somerville MA: Cascadilla.

Ozeki, Hiromi & Yasuhiro Shirai. 2007. The consequences of variation in the acquisition of relative clauses: An analysis of longitudinal production data from five Japanese children. In *Diversity in language: Perspectives and implications*, Y. Matsumoto, D.Y. Oshima, O.W. Robinson & P. Sells (Eds), 243–270. Stanford CA: CSLI.

Pawley, Andrew. 2007. Developments in the study of formulaic language since 1970: A personal view. In *Phraseology and culture in English*, P. Skandera (Ed.), Berlin: Mouton de Gruyter.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication*, J.C. Richards & R.W. Schmidt (Eds), 191–268. London: Longman.

Schachter, P. 1985. Parts-of-speech systems. In *Language typology and syntactic description*, Vol. I, T. Shopen (Ed.), 3–61. Cambridge: CUP.

Selting, Margret & Elizabeth Couper-Kuhlen (Eds), 2001. *Studies in interactional linguistics*. Amsterdam: John Benjamins.

Shibatani, Masayoshi. 1990. *The languages of Japan*. Cambridge: CUP.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: OUP.

Stefanowitsch, Anatol & Stefan T. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–43.

Thompson, Sandra A. 1988. A discourse approach to the cross-linguistic category 'adjective'. In *Explaining language universals*, John Hawkins (Ed.), 167–185. Oxford: Basil Blackwell. (Also 1989. *Linguistic categorization*, Roberta Corrigan, Fred Eckman & Michael Noonan (Eds), 245–265. Amsterdam: John Benjamins.

Thompson, Sandra A. & Elizabeth Couper-Kuhlen. 2005. The clause as a locus of grammar and interaction. *Discourse Studies* 7(4/5): 481–505.

Thompson, Sandra A. & Fangqiong Zhan. Forthcoming. Adjectives in Mandarin conversation.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge MA: Harvard University Press.

Uehara, Satoshi. 1996. Nominal adjectives in Japanese (and in Korean?). In *Japanese-Korean linguistics*, Vol 5, N. Akatsuka, S. Iwasaki & S. Strauss (Eds), 235–250.

Uehara, Satoshi. 1998. *Syntactic categories in Japanese: A cognitive and typological introduction*. Tokyo: Kurosio Publishers.

Weber, Thilo. 1997. The emergence of linguistic structure: Paul Hopper's emergent grammar hypothesis revisited. *Language Sciences* 19(2): 177–196.

Wray, Alison. 2002. *Formulaic language and the lexicon.* Cambridge: CUP.

Wray, Allison & Michael R. Perkins. 1999. The function of formulaic language: An integrated model. *Language and communication* 20: 1–28.

Wulff, Stefanie. 2007. Rethinking idiomaticity: A corpus-linguistic approach. Ph.D. dissertation, University of Bremen.

Yanabu, A. 1982. *Hon'yakugo Seiritsu Jijoo* [Circumstances for the Formation of Translationese]. Tokyo: Iwanami Shoten.

Genre-controlled constructions in written language quotatives

A case study of English quotatives from two major genres*

Jessie Sams
University of Colorado at Boulder

*I would like to thank everyone who helped me while writing this paper, including anonymous reviewers, colleagues who attended the UWM Linguistics Symposium, and my advisor Laura Michaelis. I also thank my wonderful family and friends, who have supported me every step of the way.

Abstract

Quotatives in written language serve a variety of functions, depending on genre. Newspaper quotatives often introduce new participants by providing appropriate background information and tend to be more formulaic; fiction quotatives can describe narrative-advancing events or develop characters and are often more creative. This study examines quotatives from these two major written genres and uses a set of lexical and grammatical features (quotative position, quoting verb type, nominal versus pronominal speakers, adverb and adjective use, quotative inversion, and null quotatives) to illustrate that these functional differences affect quotative form. This paper also examines quotatives from a gossip column, which is functionally and formally similar in some ways to both newspapers and fiction books, to further distinguish genre-dependent features.

1. Introduction

In written genres, quotatives are used with quotations to indicate that a quotation was spoken:

(1) a. "Hello," **Mary said.**
b. "Hey!" **yelled Mary**.

The quotative in (1a) is *Mary said* and in (1b) is *yelled Mary*; the verbs *said* and *yelled* in these examples are the quoting verbs. Both these examples are simple quotatives because they consist only of a subject and a quoting verb. Quotatives can be much more complex than this by the addition of adverbs, event sequences, or biographical information about the speaker: *Mary said as she smiled at the customer walking in the door*. All the constituents within the quotative are working together to show who is speaking and how that person is speaking; furthermore, these constituents can also have the functions of showing when the quote was spoken, what activities were happening before, during, or after the quotation, and why the speaker is qualified to make the statement in the quotation. While quotatives in spoken language are somewhat limited, there is a great deal of variability in quotatives found in written genres. A large number of verbs can be used in quotatives, including verbs that do not generally denote speech acts (*nod, smile*); also, quotatives in written language can be used in varying positions (initial, medial, final) relative to the quoted material:

(2) Initial: **As I took it out of my wallet, he continued**, "To say you were over the speed limit is putting it mildly." (Duncan 1989: 148)

(3) Medial: "I know you're new here, Lola," **purred Carla Santini**, "and you don't understand how things work yet." (Sheldon 1999: 47–48)

(4) Final: "Oh, now, Chuck," **Mrs. Bass clucked**. (von Ziegesar 2002: 203)

Along with these options, quotatives can also be null; that is, in some situations, there is no need for a quotative, and the quotation will appear without one.

By examining quotatives from written genres, I want to answer some basic questions about the nature of quotatives that have not yet been answered in previous literature, especially those pertaining to the role of the genre within which the author is writing. The quotative features I focus on in §2 are the use of the three quotative positions, quoting verbs from particular FrameNet categories, nominal versus pronominal speakers, adverbs and/or adjectives, quotative inversion, and null quotatives. While looking at these features, I will be focusing on connections between form and function; specifically, how does a genre's function interact with the form and function of the quotatives found in that genre? In §3, I examine quotative functions according to genre; in §4, I introduce a feature spectrum based on quotative features found in the data. A goal of this paper is to further the studies that have been done on the role of quotations (e.g., Clark & Gerrig 1990; Vries 2008), textual analyses of quotatives and quotations (e.g., Waugh 1995; Waugh & Monville-Burston 1986), and the syntactic relationship between quotatives and quotations (e.g., Ruppenhofer 2001; Collins & Branigan 1997) by performing a cross-genre data analysis.

1.1 Roles of quotations and quotatives in written genres

Clark & Gerrig (1990) compare quotations to gestural demonstrations. Just as you can demonstrate how someone limps by using bodily motions, you can demonstrate what another speaker said–and how she said it–by directly quoting that person. Therefore, according to Clark and Gerrig, quotations are non-serious modalities used to demonstrate a previously spoken utterance. For example, in (5), the verb *said* is a quoting verb that indicates the following speech act is the demonstration:

(5) And then she said, "Well, I'm not going to do that for you."

Because the quote is a demonstration, the recipient knows that the words in the quote are most likely not the original speaker's actual words. While this theory appears to adequately describe the functions of most quotations in spoken language, it needs to be expanded to better describe the functions of quotations and quotatives in written language.

In fiction writing, I argue that the *quotatives* are able to demonstrate of an emotional state or perceived emotional state of the speaker rather than the quotation itself being a demonstration. When an author writes quotations for characters, the words in that quotation are understood to be that character's actual words; therefore, the quotations no longer seem to be acting as demonstrations.

Quotatives can often be necessary for the reader to be able to correctly interpret the quotation:

(6) a. "You are so dead," Bridget teased like a five-year-old. (Brashares 2001: 129–30)
b. "You are so dead," Bridget shouted angrily.
c. "You are so dead," Bridget said.

In these examples, the quotatives used in (6a) and (6b) are demonstrations of the speaker's emotional state; in (6a) the speaker's quote is implying a meaning far different from the meaning implied in that same quote in (6b). The best way for an author to be sure that readers can perceive these differences correctly is through the use of demonstrative quotatives. While the first two demonstrate an emotional state, example (6c) does not; the reader is free to interpret the quotation based on the surrounding context.

In journalistic writing, such as newspapers, the quotative seems to serve a different function; rather than acting as a demonstration of the speaker's emotional state or perceived emotional state, the quotative often introduces new speakers and gives the appropriate background information about that speaker to show that he is qualified to make such a statement.

(7) "Things look very good for Ritter, but the election is still a month away, and this isn't a state where Democrats tend to win by big margins statewide," said Seth Masket, a political science professor at the University of Denver. (*Denver Post*, 10/8/06)

In (7), a previously unmentioned speaker, Seth Masket, is introduced as the speaker of the quotation. Since his name is new information, a relevant description of his background knowledge is included to assure the reader that this quotation is from a reputable source. Unlike the spoken quotations analyzed in Clark and Gerrig's study, though, these quotations are taken to be the exact words of the original speaker. Quotatives, then, can assist with the two major functions that Waugh (1995) ascribes to reported speech in journalistic writing: newsworthiness and evidentiality. That is, quotatives that provide speaker qualifications can serve to validate the assertions made in the news story.

In Waugh's (1995) article, she looks at reported speech in French newspapers in order to explore these functions that reported speech can have within journalistic writing. She states that newspapers are responsible for the information they convey through using reported speech; when using reported speech, journalists must always be aware of how to portray to their readers that their sources can be trusted to provide reliable information (129). Her paper thoroughly discusses reported speech and the variety of quoting verbs used in French newspapers but does not focus on cross-genre analyses.

1.2 Quotative constructions and formulaicity

Constructions are most simply defined as form-meaning pairs (e.g., Goldberg 2006; Kay & Fillmore 1999; Michaelis 2006; Goldberg & Jackendoff 2004). The most basic quotative construction has the form *NP QuotingVerb* and the meaning *speaker expresses directly reported speech*. Atypical quoting verbs can be coerced into a speech act reading through being used in a quotative construction by denoting the manner in which the quotation was spoken. Fauconnier & Turner (2002) talk about the fusing of verbal and constructional roles as "conceptual blending;" they state that constructions allow a speaker to compress an event and express that event in language through constructions, which have "a stable syntactic pattern that prompts for a specific blending scheme" (369). The construction supplies a way of talking about an event, and the event supplies the specific words and ideas.

In saying that the quoting verb often gives information about the manner in which the quotation was spoken, I do not mean to imply that all quoting verbs indicate the same type of manner. For instance, some quoting verbs purely indicate the manner of speech: *shout, yell, whisper*; other quoting verbs, however, give information about the emotional stance of the speaker toward a specific person or topic: *accuse, praise*; others provide information about the speaker's affect and manner of speaking: *coo, cluck, singsong*; and yet others are neutral in terms of manner: *say, tell, comment*. Examples of creative quoting verb use are in (8):

(8) a. "Regretfully, I have seen some parents taking pictures, asking for autographs, talking to the media and even shouting at Ms. Jolie and Mr. Pitt for recognition," he **finger waggles**. (*Hot Gossip*, 9/27/07)

b. "I don't know, to be honest with you. I haven't spoken to her in years," **sidestepped** the chart-topper, who is currently dating the baggage-free Jessica Biel. (*Hot Gossip*, 9/27/07)

c. "Siddown," she **Don Corleoned**. "Okay, the good news is that you're not fired. Yet." (Keyes 2006: 321)

By using these verbs in the quotative construction, a reading of saying is coerced through a relation of manner.

Wray & Perkins (2000: 1) provide the following definition for a formulaic sequence:

> a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

This definition aligns with others provided by such authors as Coulmas (1979), Overstreet & Yule (2001), and Fillmore (1977). An important feature of formulaic language is predictability, whether specific words or grammatical sequences are predictable. As will be seen in §2, quotatives have specific components that have a similar underlying function of showing who said what. Wray (1999) examines the forms and functions of formulaic sequences and states that formulaic sequences are more common in speech than in writing but that writers can "employ formulaic sequences in their writing as a stylistic device, particularly to indicate the discourse status" (227–228). At the discourse level, quotatives help keep track of turn-taking sequences (who has the current turn) but do not necessarily have to occur at the boundaries of the turn. Medial quotatives always occur inside an utterance, and initial and final quotatives can be used within a speaker's turn. Quotatives have varying degrees of formulaicity; one aim of this paper is to show that genre dictates the degree of quotative formulaicity, both in specific lexical choices and constructional patterns.

1.3 Genre effects and data sources

Based on the previously mentioned studies and my own intuitions about language use, I chose to compare two genres that provide data on opposite ends of a quotative-use spectrum based on the genre's function. Young adult fiction books provide one end of a written-genre spectrum, with a large variety of quoting verbs and of available patterns for the quotatives. In total, I used twenty young adult fiction books for collecting data from fiction sources. Newspapers provide another end of the spectrum, with less creativity in quoting verb selection and more restrictions in available patterns for the quotatives. For collecting data from newspapers, I used ten newspapers from across the country; in order to avoid 'sub-genre' effects, I collected data from both printed and online newspapers, which represented a variety of geographical locations and distribution ranges, from a number of dates between October 2006 and March 2007; and from all types of articles within the newspapers (e.g., front page, business section, technology, entertainment).

In my data, there are more instances of quotatives for fiction books than there are for newspapers. The reason for this is that as I was collecting data, I was looking for any emerging patterns within quotative use. While these patterns stabilized within the first 200 tokens of newspaper quotatives, the patterns in fiction book quotatives remained opaque. I also wanted a variety of author's styles for a more representative sample; while newspapers often showcase the writing of several authors, fiction books tend to only have one author. For these reasons, I believe that even though there are fewer instances within the newspaper genre, the sampling is just as representative of the population at large as the data for fiction books.

In order to more fully explore the effect of genre on quotative use, I collected a small sample of data from a genre that falls between newspapers and fiction books: MSN.com's *Hot Gossip* by Kat Giantis. A major function of the gossip column is similar to that of a newspaper's function: to inform the reader and give a variety of perspectives on a given event. As such, reported speech used in gossip columns is still perceived to be the actual words of the original speaker. Due to this similarity in function, one would expect the same features found in the quotatives of newspapers to be found in the quotatives of the gossip column. Unlike newspapers, though, *Hot Gossip* is also written like a storyline that might be included in a fiction book and is not written to be unbiased. I argue that this dual function affects quotative use.

1.4 Annotating quotatives

The features I annotated for every quotative are quoting verb type (individual quoting verbs were recorded and then classified according to FrameNet; see § 2.2.1), quotative position (initial, medial, final), morphological expression of the speaker role (nominal [X] or pronominal [p]), and inversion (e.g., *asked Ella*). Also, the number of times quotations appeared without a quotative was recorded (null quotative).

In order to be a null quotative, a quotation has to appear without a quotative; while this seems to be straightforward, final quotatives are often followed by continuing dialogue from the same character:

(9) "No I'm not! I'm mad at those other people!" Carmen shot back. "I don't want to have anything to do with them. I want them to go away and for it just to be me and my dad again." (Brashares 2001: 197)

In (9), there is one final quotative: *Carmen shot back*. Even though the quotation continues after the use of the quotative, I did not count the second half of the quote as having a null quotative because the addressee and manner of speech are both the same. To use another quotative for the second half of the quote would have been redundant and unnecessary. An example of a continuing quotation that changes addressee and manner follows:

(10) "*My* friends. . . *my* friends. . . ," he chanted as we dragged him along. "*My* friends . . . we're going for a drink. . . ." And then he made one of his sudden stops. "Who are you?" He was shouting again. "You're not my friends. I don't have any friends." He started laughing again. "Not unless they want something from me. What do you want?" (Sheldon 1999: 234)

In this paragraph, it is a little more difficult to count the quotatives. Halfway through the quotation, the addressee changes because the speaker goes

from chanting, which implies there is not a particular addressee, to directing his attention to the girls walking with him. Furthermore, his manner changes from the beginning of the quotation to the end. Therefore, in this paragraph, there is one final quotative (*he chanted as we dragged him along*) and one null quotative.

2. Quotatives in written English

The aim of this section is to align quotative variability with functional variability; in other words, what are the different forms quotatives can take? Do these forms correlate with specific functions? This section also explores the specific effects of genre on quotative use. For some of the questions, I needed to take a random sampling in order to better analyze the data. The sampling for fiction books includes 120 instances of quotatives; for each position (initial, medial, final), there are 40 examples. The sampling for newspapers includes 100 instances of quotatives; there are 35 examples for initial quotatives, 20 examples of medial quotatives, and 45 examples of final quotatives. The gossip column sampling includes 60 instances of quotatives; for each position, there are 20 quotatives. The random samplings were mainly used for a qualitative analysis (especially in terms of the use of adverbs and adjectives) instead of a quantitative analysis.

2.1 Quotative positions

As stated in §1, quotatives can appear in the initial, medial, or final positions relative to the quotation. If no meaning were placed on the position of the quotative, one of two patterns would emerge: either there would be only one position, or the positions used would be nearly equally distributed. However, the data do not support this; in Table 1, both the raw numbers and percentages are provided:

Table 1. Positions by genre

	Initial	Medial	Final	Total
Newspapers	227/16.67%	29/2.13%	1106/81.20%	1362
Fiction books	846/5.92%	350/2.45%	13102/91.63%	14299
Gossip column	127/54.04%	5/2.13%	103/43.83%	235

As Table 1 demonstrates, there is more than one position utilized, and the three positions are not equally distributed in the data.

The percentages of medial quotatives is nearly equal for all genres and is very low. Also, both newspapers and fiction books most heavily rely on the final position, which appears to be a default position for these genres; however, the percentage of initial quotatives is higher in newspapers than in fiction books. When initial quotatives are used in fiction books, there is often an emphasis placed on the integration of the storyline into the quotative; consequently, many initial quotatives include adverbial clauses or a sequence of events:

(11) A few minutes later, when we were back on the freeway, Lorelei suddenly said, "You may have been right about that car. It should have passed us while we were stopped, but it didn't." (Duncan 1989: 148–149)

(12) She laid the material on her chair, turned to Ben, and said, "Let's dance." (Dean 2003: 99)

In (11), the quotative includes the adverbial clause *when we were back on the freeway*; in (12), the quotative includes a sequence of events: laying the material on the chair, turning to Ben, and saying her quotation. However, this role is not a necessary one, as any given sequence of events can take place outside the quotative and does not need to be integrated into the quotative itself:

(12′) She laid the material on her chair and turned to Ben. "Let's dance," she said.

This rewording of (12) portrays the same information in a different manner. Although initial quotatives are useful to help further the storyline, they are not necessary to do so.

In newspapers, initial quotatives are often used to introduce partial quotations (a mixture of direct and indirect speech), as in (13):

(13) Berk said his client "wanted to do the right thing" and agreed that there was some back child support owed from before his imprisonment. . . (*Daily Report*, 3/2/07)

This pattern is often utilized in journalistic writing when the author of an article is working with incomplete quotations, material where only part of the quotation is usable for the article, or the integration of many different quotations from the same source (Waugh 1995). This mix of direct and indirect speech is often necessary for the author to work quotations into the overall story. Because mixing direct and indirect speech is a necessary feature of journalistic writing and the initial position is the most common position that works with this feature, the initial position is utilized quite often in journalistic writing.

The gossip column differs from both newspapers and fiction books in that the percentage of initial quotatives is higher than final quotatives. This is because, like

newspapers, the gossip column utilizes partial quotes; however, the gossip column relies on these much more frequently than newspapers and, therefore, uses the initial position more frequently:

(14) Last week, Tom's ex, Penelope Cruz, gushed that the infant was "really beautiful–she's one of the most beautiful babies I've ever seen," before quickly clamming up on the subject. (*Hot Gossip*, 8/9/06)

Many examples like (14) can be found in *Hot Gossip*. Kat Giantis freely quotes sources and intertwines those quotations into her own personal diatribe to write a story that flows more like a fictional short story. Reliance on the initial position is one of the defining features of quotative use within written genres.

2.2 Forms of quotatives

In the data, five major quotative components emerged:

(15) Possible components of a quotative
- a. quoting verb
- b. speaker
- c. adverb*
- d. adjective*
- e. addressee(*)

The components with an * are optional; that is, adverbs, adjectives, and sometimes the addressee are not required for grammaticality while quoting verbs and speakers are.

2.2.1 *Quoting verbs*

The most common quoting verb in the data is *say*; the next four most common quoting verbs are *ask, tell, whisper,* and *call*. Within quoting verb selection, there is a wide variety of verbs that can be used that do not typically denote a speech act, such as *explode, yawn, frown, ooze*, and *thunder*. In order to discuss the types of verbs allowed, I used a FrameNet (http://framenet.icsi.berkeley.edu) categorization of the quoting verbs, which is also used by Ruppenhofer (2001). In Ruppenhofer et al. (2006), the following definition is provided for a semantic frame:

> [It is] a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props. For example, the Apply_heat frame describes a common situation involving a COOK, some FOOD, and a HEATING_INSTRUMENT, and is **evoked** by words such as *bake, blanch, boil, broil, brown, simmer, steam,* etc. (5)

An example chart of the FrameNet classification is in Table 2; this chart includes some of the frames utilized by quoting verbs and examples of quoting verbs within the given frames from my corpora:

Table 2. FrameNet Classifications of Quoting Verbs

Frame	Examples of Quoting Verbs
Body/Body-movement	clap, shrug, yawn
Body/Making-faces	frown, grin, sneer
Body/Respiration	breathe, gasp, sigh
Cognition/Becoming-aware	note, observe
Cognition/Judgment	boast, compliment, scold
Communication/Manner	babble, mutter, shout, slur, stammer, whisper
Communication/Noise	bark, call, cry, hiss, scream, snap, yell
Communication/Question	ask, inquire
Communication/Statement	announce, comment, declare, explain, inform, remark, say, state, tell
Emotion/Pressure	blurt (out), explode, gush, ooze
Perception/Experience	hear, overhear
Suasion/Attempt-suasion	advise, encourage, urge, warn

For each genre and each larger frame category (i.e., *Body* encompasses all the specific Body frames, such as Body/Making-faces and Body/Respiration), Table 2.1 shows the number of quoting verbs that fit into that particular frame (QVs), the number of instances quoting verbs from that frame were actually utilized in the data (Inst), and the percentage of quoting verb use within each frame (%).

Table 2.1. Quoting verb use by frame and genre

	Newspapers			Fiction books			Gossip column		
Frame	**QVs**	**Inst**	**%**	**QVs**	**Inst**	**%**	**QVs**	**Inst**	**%**
Body	0	0	0	25	116	0.81	1	4	1.70
Cognition	5	9	0.66	47	412	2.88	8	9	3.83
Communication	33	1348	98.97	231	13554	94.76	53	215	91.49
Emotion	1	1	0.07	6	40	0.28	1	4	1.70
Perception	0	0	0	2	54	0.38	2	2	0.85
Suasion	3	4	0.29	14	127	0.89	1	1	0.43

For instance, newspapers utilize 5 different quoting verbs from the Cognition frames; between those 5 quoting verbs, there are 9 instances of Cognition-frame quoting verbs in the data. Given the low number of instances (9), the percentage of Cognition frame quoting verb use in newspapers is also very low (0.66%). Of the

six frames, quoting verbs found in the Communication frames are most prevalent in all genres, which portrays the sense that quoting verbs often belong to the same frame; however, most quoting verbs in newspapers specifically belong to the Communication/Statement frame, as shown in Table 2.2.

Table 2.2. Communication/Statement frame quoting verbs

	Newspapers			Fiction books			Gossip column		
	QVs	Inst	%	QVs	Inst	%	QVs	Inst	%
Comm/Statement	17	1289	94.64	48	8936	62.48	27	161	68.51

The percentage of quoting verbs belonging to the Communication/Statement frame is much lower for fiction books (62.48%) and the gossip column (68.51%) than for newspapers (94.64%). This shows that while the Communication frames more frequently occur in all the genres, the fiction books and gossip column have more variety within the Communication frames.

Over 87% of the quoting verbs in the newspaper data are the verb *say*; if other quoting verbs are used, another quoting verb from the Communication/Statement frame is most likely utilized:

(16) "SoHi shoots the ball very well. We talked about it at halftime, getting out on shooters. But again, it's inexperience," Skyview coach Red Goodwin **explained**. (*Peninsula Clarion*, 3/4/07)

In this example, and others like it, *say* is not the quoting verb; however, the quoting verb *explain* also belongs to the Communication/Statement frame.

Quoting verbs in the fiction books and gossip column have much more lexical variety– even when only looking at quoting verbs within the Communication frames–than those found in newspapers:

(17) "Oh, *no*," she **moaned** in her best mommy voice. (Pascal 2004: 128)

(18) "I was saying," Cosgrove **went on**, still glaring, "that my attorney has provided you with files on both Jaffarian and Sullivan, and– " (Golden & Hautala 2004: 180)

The quoting verb in (17), *moaned*, belongs to the Communication/Noise frame while the quoting verb in (18), *went on*, belongs to the Communication/Turn-taking frame. These are both examples of quoting verbs that belong to the Communication frames but do not belong to the Communication/Statement frame.

Hot Gossip has more flexibility in quoting verb selection than the other genres, as in (19):

(19) "What's funny about that is that Wilmer and I have never dated," the "Ghost Whisperer" starlet **pooh-poohed** last week during an interview with a CBS affiliate in Sacramento, Calif. (*Hot Gossip*, 4/13/06)

Other quoting verbs only found in the gossip column data include *tattled, warbled, snitched,* and *overshared*. As could be expected in a gossip column, the frame Communication/Reveal-secret is utilized more often than in either fiction books or newspapers. The reliance on the Communication/Statement frame for quoting verb selection is another feature that separates written genres.

2.2.2 *Speaker*

The speaker of the quoting verb appears in one of two forms: nominal or pronominal. The distribution of these types of speakers in the written genres is shown in Table 3; "X" stands for a nominal speaker, and "P" stands for a pronominal speaker:

Table 3. Speaker role by genre

	X	P
Newspapers	1051/77.17%	311/22.83%
Fiction books	7361/51.48%	6938/48.52%
Gossip column	193/82.13%	42/17.87%

As this table demonstrates, the distribution depends on the genre; nominal speakers are heavily favored in the newspapers and gossip column while the distribution between nominal and pronominal speakers is nearly equal in fiction books. The fact that pronouns are not used as often in newspapers is a predictable feature of journalistic writing, as it is the author's job to make sure the reader knows exactly who is saying what (Waugh 1995).

Because pronouns can lead to confusion if more than one speaker has been mentioned, newspapers are more likely to use nominal speakers. Because the gossip column attributes quotations to specific sources (like newspapers), the speaker is more likely to be nominal to avoid any possible confusion as to the identity of the source. Within the fiction books corpus, some books are written in the first person; books written in the first person show a higher tendency to use pronominal speakers than other fiction books. The use of pronominal speakers, therefore, is genre-dependent.

2.2.3 *Adverbs and adjectives*

Adverbs and adjectives are optional elements of quotatives; these include single-word adverbs and adjectives, as well as phrases and clauses. The data in Table 4 is taken from the random samplings of each genre:

Table 4. Adverb and adjective use by genre

	Adverbs	Adjectives
Newspapers	8/8.0%	37/37.0%
Fiction books	62/51.7%	3/2.5%
Gossip column	7/11.7%	9/15.0%

Adverbs in quotatives are generally used to show how or when a particular quotation was spoken and are more common in fiction book quotatives. Adjectives, on the other hand, are used more often in newspaper quotatives in order to give background information for the speaker. The gossip column, however, does not exhibit high uses of either adverbs or adjectives and so in this case aligns with newspapers in low adverb use and with fiction books in low adjective use. One difference between the use of adverbs in the gossip column quotatives and the other genres is that when an adverb is used, it often demonstrates the *author's* emotional stance toward the quoted speech rather than the speaker's:

(20) "… [S]ometimes what I actually love to do is go to a farm and get fresh milk or watch a pig get slaughtered," he ookily explains. (*Hot Gossip*, 1/26/07)

In this example, the author is letting the readers know how she feels about this particular quotation by using the adverb *ookily* within the quotative.

2.2.4 *Addressee*

The addressee component is sometimes required and sometimes not, depending on the quoting verb but not the genre. For instance, verbs like *tell* require an addressee:

(21) a. "If it's a hacker," she told him, "it's a criminal offense." (McDonald 2001: 20)
b. * she told

However, verbs like *ask* do not require an addressee:

(22) a. "Blair doesn't know, does she?" Serena asked Nate quietly. (von Ziegesar 2002: 35)
b. Serena asked quietly

Even though the addressee is absent in (22b), it is still grammatical. Often, if the addressee is optional, it will appear in a prepositional phrase:

(23) "This was your *life*," he shouted to the empty shell overhead. (McDonald 2001: 219)

Shout is a quoting verb that does not require an addressee, and in (23) the preposition *to* is used to indicate the addressee. While the addressee is a

constituent within quotatives, it does not appear to be a genre-dependent feature in the data.

2.3 Quotative inversion

In quotative inversion, the quoting verb precedes the speaker, as in (24):

(24) "Have you listened to one word I've said?" asked Ella. (Sheldon 1999: 96)

Typically, quotative inversion does not occur in modern English written sources if the speaker is pronominal; in fact, there is only one example of this in all the data:

(25) "Nothing ventured, nothing gained," said I. (Sheldon 1999: 94)

Also, the ability for quotative inversion to appear in the initial position depends on the genre; while fictional sources do not show any instances of quotative inversion in the initial position, newspaper sources do utilize this pattern:

(26) Said Elway: "I think the two locations will draw very different crowds."
(*Denver Post*, 1/26/07)

Quotative inversion also depends on the quoting verb selection; quoting verbs that require an addressee (or that have an addressee without a preposition) cannot appear in quotative inversion:

(27) a. "Well. . . the staring part isn't so bad," Jenny told her, "but the drooling is not very subtle. I asked you about Eugene." (Stine 1989: 38)
b. *told Jenny her
c. *told her Jenny

(28) a. "Tell me you don't care about Serena van der Woodsen being back," Jenny challenged Dan. (von Ziegesar 2002: 92)
b. * challenged Jenny Dan
c. * challenged Dan Jenny
d. challenged Jenny

(29) a. "She did all our hymnals at school," Kati whispered to Tina.
(von Ziegesar 2002: 126)
b. whispered Kati to Tina

In (27), the quoting verb *tell* requires an addressee and cannot be inverted. In (28), the quoting verb *challenge* does not require an addressee but does not use a prepositional phrase to include an addressee; thus, when the addressee is present, *challenge* cannot appear in quotative inversion, as in (28b) and (28c). If the addressee is deleted, though, *challenge* can appear in quotative inversion, as in (28d). In (29), the quoting verb *whisper* does not require an addressee and places the addressee

in a prepositional phrase; even when the addressee is present, *whisper* can appear in quotative inversion, as in (29b).

Two-word quoting verbs can also affect the possibility of quotative inversion:

(30) a. "She's a sweet, sweet kid, your friend," Angela went on. (Brashares 2001: 272)

b. ? went on Angela

The use of the quoting verb *went on* in an inverted quotative sounds awkward and does not appear in the data; however, there are several other two-word quoting verbs that do appear inverted in the data, as in (31):

(31) "You know," chipped in Carla, "she could be Korean." (Sheldon 1999: 67)

Chipped in, chimed in, and *put in* were the three two-word quotatives that appear inverted in the data; the fact that they were the only ones inverted may be a function of the second word of the two-word verb being *in*.

The use of quotative inversion also differs by genre:

Table 5. Quotative inversion by genre

	Inverted Quotatives	**Total Quotatives**
Newspapers	326/23.94%	1362
Fiction books	1055/7.38%	14303
Gossip column	33/14.04%	235

Inverted quotatives are three times as likely to appear in newspapers than in fiction books. While this is the case, newspapers are more restricted with the quoting verb that is used in the inverted construction. Out of all the instances of inversion within newspapers, only four of these are instances with a quoting verb other than *say*; these four instances are with four different quoting verbs: *ask, complain, explain,* and *write*. In the fiction books and gossip column, on the other hand, while *say* is also the most common quoting verb used in the inverted construction, many other quoting verbs are utilized in inverted quotatives, such as *add, beg, call, cry, hiss, mutter,* and *whisper*. Within fiction writing, it is evident from the data that the use of inversion is also dependent on the preference of the individual author; of all the cases of inversion within the fiction data, nearly 20% of these instances are attributed to a single author, Jacqueline Wilson.

Inverted quotatives are often utilized in newspapers when the speaker is heavy:

(32) "Most of this curriculum is really poor, with no health message," said Mary McCourt, a community health specialist at the Missoula City-County Health Department. (*Missoulian*, 3/4/07)

As in (32), when background information is given about the speaker, the quotative often appears as an inverted quotative. Because the gossip column does not heavily rely on the practice of giving that background information about the speaker, the number of instances of inverted quotatives is much lower than the number in newspapers. The use of inverted quotatives is also a feature that determines genre classification.

2.4 Null quotatives

There is a preference for using a quotative when one could be used; this shows that even though the authors have a choice in how to make it clear who is speaking, they favor the use of quotatives to achieve this goal. In Table 6, the *Total* column refers to the number of instances in which a quotative could have been used; in other words, it is the total number of quotatives actually used added to the number of null quotatives.

Table 6. Null quotative use

	Null quotatives	Total of possible uses
Newspapers	27/1.94%	1389
Fiction books	9152/39.02%	23455
Gossip column	27/10.31%	262

As stated earlier, one job a journalist must accomplish is to make it explicitly clear who is saying what; therefore, it follows that newspapers have a very low occurrence of null quotatives. In fact, they will often be redundant in asserting the same source time after time without a change of speaker to be assured that there will be no confusion on the reader's part as to who originally stated the quotation. In fiction books, however, null quotatives can be used more often. *Hot Gossip* also does not have a high number of instances of null quotatives. This can be attributed to the similarity in function of gossip columns and newspapers: When there are quotations, it is the author's job to make sure readers are not confused as to who said what. Therefore, newspapers and gossip columns tend to always state who the speaker is, rather than relying on the reader to be able to conclude who the speaker is based on surrounding information; the ability to utilize null quotatives is another genre-dependent feature.

3. Functions of quotatives

The goal of this section is to identify what can be accomplished through the use of quotatives in each genre.

3.1 Newspapers

In newspapers, the emphasis is placed on the quotation; a quotative is generally used as a tool to show who the source is and why the reader can trust that source. Moreover, the speaker is generally introduced to the readers in a quotative. Referent introduction is most often achieved through the addition of an adjective clause to give some background for a particular speaker.

(33) "The brown pelican, the American alligator and the peregrine falcon are prime examples of recovered species that now occupy nearly all of their historic range," said Rob Edward, **director of carnivore restoration for Sinapu, a Colorado-based advocacy group for wolves and predators**. (*Denver Post*, 1/30/07)

In (33), the quotative includes a rather detailed adjectival phrase used to describe the speaker, Rob Edward. The quotative tells the reader that Mr. Edward is the *director of carnivore restoration for Sinapu*; while this is a full description, the journalist could not be guaranteed that all readers would know the group *Sinapu*. Therefore, a description for Sinapu was also added: *a Colorado-based advocacy group for wolves and predators*. By using these descriptors, the reader now knows why the author believes this is a trustworthy source. As Waugh (1995: 132) states,

> Readers, even average, naive readers, are likely to be skeptical about the well-foundedness of the facts that are being presented to them and need to be persuaded that indeed they are facts, especially since they know that the average reporter cannot know everything and cannot be sure of everything. So, while writing their reports, reporters . . . turn to experts for verification of specific points . . . .

As Waugh points out, journalists need to use experts' opinions (as well as eye witness accounts) to write a well-rounded story; because readers are not likely to recognize all the names used as sources, it is the journalist's job to include pertinent background information about the source that will identify that source as a reliable source. In newspapers, this information is often integrated into the quotative, as in (33).

3.2 Fiction books

In fiction, there is an emphasis on the storyline; as such, quotatives are often woven into the storyline and can be used to further this storyline by adding adverbial clauses to tell the reader what is going on during the speech act:

(34) a. "Anyway," Christina continued **while drying her face on a paper towel**, "Bill offered to drive me home. . ." (McDaniel 2002: 3)

b. "Here. You," Jacqui said, **pushing the Pringles toward Madison**, "What is your name?" (de la Cruz 2004: 63)

A quotative can also further the storyline when the quoting verb is placed in a sequence of events:

(35) "That should cheer her up, don't you think?" **she whispered and glanced over at the woman sleeping in the hospital bed**. (McDonald 2001: 136)

Quotatives can also be used to demonstrate the speaker's emotional state and thus contribute to the function of character development. Character development is most often achieved through the use of quoting verb selection (36) or the addition of adverbs to show a character's particular stance or emotional affect (37).

(36) a. "Jimmy," I **quaver**, "when I said there's a storm coming, I meant, there's a storm coming! I have a leaky sunroof! I never told you to burn your club!" (Korman 2002: 230)
b. "You should have seen him, Dad," I **plead**. "He's terrified! He thinks Uncle Shank is going to cut off his fingers." (Korman 2002: 109)

(37) a. "Well, Cody isn't any threat, is he? I mean, the accident–*your* accident," she added **hotly**, "wiped his mind clean. That must have been a real plus for you." (McDaniel 2002: 210)
b. "Hey, don't you have to go burp my brother or something?" she said **snidely**. (de la Cruz 2004: 222)

3.3 Gossip column

As stated earlier, the dual function of the gossip column affects quotative usage by making the quotatives in the gossip column similar to fiction books in some ways but similar to newspapers in other ways. The gossip column does not often integrate the storyline into the quotatives but rather weaves a storyline through the integration of indirect and direct speech; also, it relies on information from sources that prefer to remain anonymous (e.g., *a fly-on-the-wall snitch*–sources like these allow the readers to make their own judgments as to whether or not they should believe the quotation) or are celebrities who need no introduction. Therefore, the gossip column does not often utilize the functions of speaker introduction or furthering the storyline. However, quotatives in the gossip column can be used to show the emotional stance of either the author or the speaker and can lend to the function of speaker development in the same way that fiction book quotatives can develop characters: through quoting verb selection (38) and the addition of adverbs (39).

(38) "She would only ever have one but I always had to fetch her both," he **kvetches**. (*Hot Gossip*, 9/24/07)

(39) "She's just cool," he **nonspecifically** enthuses of his fiancée. (*Hot Gossip*, 4/13/06)

Hot Gossip uses creative quotatives to increase its entertainment value rather than its informative value.

4. Feature spectrum

In order to visually compare genres according to their unique features and functions (discussed in the sections above), Figure 1 compares the genres' uses of null quotatives (Null), adverbs (Adv), pronominal speakers (Pro), quoting verbs within the Communication/Statement frame (C/S), initial position (Init), inverted quotatives (Inv), and adjectives (Adj).

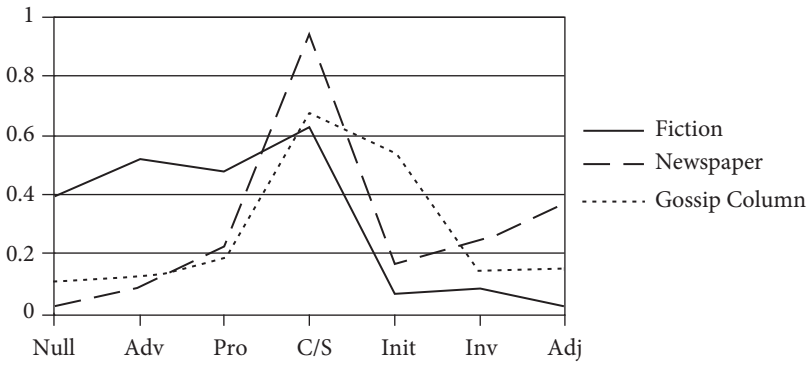


Figure 1. Comparing quotative use across genres.

The features on the left are more typical for fiction quotatives; moving to the right brings a graded shift from fiction writing to journalistic writing. The features on the right are more typical for newspaper quotatives. On the one end, the gossip column more closely follows quotative use in newspapers; on the other end, it more closely follows quotative use in fiction books (with the exception of the high number of initial quotatives, which does not follow either genre).

Not only does moving across this spectrum change the features used, it also brings a shift in the degree of formulaicity of quotatives. Journalistic writing more often utilizes predictable, formulaic quotatives while fiction writing often has a greater degree of creativity in quotatives. This is unsurprising, given the functions of the genres. Because newspapers are meant to inform readers and not necessarily entertain them, the greatest focus is on accurately presenting information. Within newspapers, 55% of all quotatives used are *X/p said* (*X* being a nominal speaker, and *p* being a pronominal speaker); another 20% of all quotatives are *said X, DESC*

(where *DESC* denotes the addition of background information for the speaker). This is even more distinct when newspaper quotatives are separated into classes: quotatives that introduce a new speaker and quotatives that feature an already active speaker. Over 60% of the quotatives that introduce a new speaker are *said X, DESC*; once a speaker has been introduced and is active within the discourse, 83% of the quotatives are *X/p said*. Therefore, newspaper quotatives are more predictable in quoting verb choice and in constructional pattern choice.

Quotatives in fiction books are more creative than those in newspapers; roughly 45% of the quotatives in the fiction book sampling are *X/p QuotingVerb* with flexibility in quoting verb selection and in the addition and placement of adverbs. Specific patterns of quotatives are dependent on the individual author's style. Newspaper quotatives are not generally dependent on authors' styles; newspapers focus on the information gathered and given rather than on style. Therefore, newspaper stories are often written in such a way that authorial style varies little from author to author. The gossip column is the most creative in terms of quoting verb selection but is more restricted in terms of the addition of adverbs or adjectives; nearly 73% of quotatives in the sampling are also *X/p QuotingVerb*.

5. Conclusion

Quotatives in written English are greatly affected by the genre of that written source. The two major genres from this study are fiction books and newspapers. While analyzing the data, genre-dependent patterns emerged within the use of specific quotative features. The spectrum of features in Figure 1 shows that fiction books are more likely to depend on the use of null quotatives, adverbs or adverbial phrases or clauses, and pronominal speakers. Newspapers, on the other hand, are more likely to depend on quoting verbs in the Communication/Statement frame, initial quotatives, inverted quotatives, and adjectival phrases or clauses. The dependence on these features closely relates to the function of each of the genres. Therefore, the controlling feature for how quotatives are used–which also affects the syntactic form of the quotative–is the genre classification of the written source.

In order to find out if deviating from the expected features for quotatives within a genre changes the perceived genre classification, data was collected and analyzed from MSN.com's *Hot Gossip* by Kat Giantis. A major function of the gossip column is similar to that of a newspaper's function: to inform the reader and give a variety of perspectives on a given event. Due to this similarity in function, one would expect the same features found in the quotatives of newspapers to be found in the quotatives of the gossip column. Unlike newspapers, though, *Hot Gossip* is also written like a storyline that might be included in a fiction

book and is not written to be unbiased. This dual function affects the usage of quotatives. By relying on sets of features expected of both newspapers and fiction books, the gossip column achieves the dual stylistic function of informing while entertaining the readers.

Data sources

Fiction Books:

Brashares, Ann. 2001. *The Sisterhood of the Traveling Pants*. New York: Random House.
Cabot, Meg. 2002. *All-American Girl*. New York: HarperCollins Publishers.
Caletti, Deb. 2002. *The Queen of Everything*. New York: Simon and Schuster.
Clark, Catherine. 2003. *Frozen Rodeo*. New York: HarperCollins Publishers.
Clark, Catherine. 2004. *Maine Squeeze*. New York: HarperCollins Publishers.
Cooney, Caroline B. 1997. *The Terrorist*. New York: Scholastic Inc.
Cormier, Robert. 1995. *In the Middle of the Night*. New York: Bantam Doubleday Dell Publishing Group.
de la Cruz, Melissa. 2004. *The Au Pairs*. New York: Simon and Schuster.
Dean, Zoey. 2003. *The A-List*. New York: Little, Brown and Company.
Donnelly, Jennifer. 2003. *A Northern Light*. Orlando: Harcourt Inc.
Duncan, Lois. 1989. *Don't Look Behind You*. New York: Bantam Doubleday Dell Publishing Group.
Golden, Christopher and Rick Hautala. 2004. *Last Breath*. New York: Simon and Schuster.
Korman, Gordon. 2002. *Son of the Mob*. New York: Hyperion Paperbacks.
McDaniel, Lurlene. 2002. *Telling Christina Goodbye*. New York: Random House.
McDonald, Joyce. 2001. *Shades of Simon Gray*. New York: Random House.
Pascal, Francine. 2004. *Gone (Fearless #36)*. New York: Simon and Schuster.
Sheldon, Dyan. 1999. *Confessions of a Teenage Drama Queen*. Cambridge: Candlewick Press.
Stine, R.L. 1989. *The Baby-Sitter*. New York: Scholastic Inc.
Wilson, Jacqueline. 1997. *Girls in Love*. New York: Random House.
von Ziegesar, Cecily. 2002. *Gossip Girl*. New York: Warner Books.

Newspapers:
Printed

Denver Post. Oct 2006 – Mar 2007.
New York Times. Oct 2006 – Mar 2007.
Peninsula Clarion. (Kenai, Alaska) Oct 2006 – Mar 2007.
St. Louis Post-Dispatch. Oct 2006 – Mar 2007.
USA Today. Oct 2006 – Mar 2007.

Online

Athens Review. (Athens, Texas) Oct 2006 – Mar 2007. ⟨http://www.athensreview.com⟩.
CNN.com. Oct 2006 – Mar 2007. ⟨http://www.cnn.com⟩.
Daily Report. (Atlanta, Georgia) Oct 2006 – Mar 2007. ⟨http://www.dailyreportonline.com⟩.
Los Angeles Times. Oct 2006 – Mar 2007. ⟨http://www.latimes.com⟩.
The Missoulian. (Missoula, Montana) Oct 2006 – Mar 2007. ⟨http://www.missoulian.com⟩.

Gossip Column:

Giantis, Kat. *Hot Gossip*. Apr 2006-Mar 2007. ⟨http://www.msn.com⟩.

References

Clark, Herbert H. & Richard J. Gerrig. 1990. Quotations as demonstrations. *Language* 66: 764–805.

Collins, Chris & Philip Branigan. 1997. Quotative inversion. *Natural Language and Linguistic Theory* 15: 1–41.

Coulmas, Florian. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3: 239–266.

Fauconnier, Gilles & Mark Turner. 2002. *The way we think: Conceptual blending and the mind's hidden complexities*. New York NY: Basic Books.

Fillmore, Charles J. 1977. The need for a frame semantics within linguistics. In *Statistical methods in linguistics*; H. Kalgren (Ed.), 5–29. Stockholm: Scriptor.

Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford: OUP.

Goldberg, Adele & Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language* 80: 532–568.

Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The what's X doing Y? construction. *Language* 75(1): 1–33.

Keyes, Marian. 2006. *Anybody out there?* New York NY: HarpersCollins.

Michaelis, Laura A. 2006. Construction grammar. *The encyclopedia of language and linguistics,* 2nd Edn. Vol. 3, K. Brown (Ed.), 73–84. Oxford: Elsevier.

Overstreet, Maryann & George Yule. 2001. Formulaic disclaimers. *Journal of Pragmatics* 33: 45–60.

Ruppenhofer, Josef. 2001. Direct speech reporting clauses as parentheticals. Ms.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson & Jan Scheffczyk. 2006. *FrameNet II: Extended theory and practice*. (Online at: http://framenet.icsi.berkeley.edu/book/book.pdf)

Vries, Mark de. 2008. The representation of language within language: A syntactico-pragmatic typology of direct speech. *Studia Linguistica* 62(1)

Waugh, Linda. 1995. Reported speech in journalistic discourse: The relation of function and text. *Text* 15(1): 129–173.

Waugh, Linda R. & Monique Monville-Burston. 1986. Aspect and discourse function: The French simple past in newspaper usage. *Language* 62: 846–878.

Wray, Alison. 1999. Formulaic language in learners and native speakers. *Language Teaching* 32(4): 213–231.

Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20: 1–28.

Some remarks on the evaluative connotations of toponymic idioms in a contrastive perspective

Joanna Szerszunowicz
Uniwersytet w Białymstoku

Abstract

The focal issue is the axiology of Polish and Italian toponymic idioms analyzed in a contrastive perspective. The study is a continuation of the analysis of the evaluative effect of selected idiom components. So far, the author has analyzed the axiology of Polish and Italian faunal idioms. The axiologically marked idioms convey evaluations in various ways; i.e., the core evaluative element is the toponym, the evaluation results from the meaning of the whole unit, and the etymological reinterpretation of the toponym is necessary for decoding the evaluation. The relation between language and culture is important since

toponymic components are culture-bound units of universal, national or local character, which influences the translatability of idioms.

Keywords: idiom; toponym; evaluative connotation; equivalence

1. The axiology of toponymic idioms

Idiomacity is a universal phenomenon in natural languages, which is best defined by multiple criteria. Therefore, a set of factors should be taken into consideration in order to classify a string as an idiom. According to Moon (1998: 6–8), idioms are units of formulaic character since they are characterized by lexicogrammatical fixedness, i.e., formal rigidity, which implies some degree of lexicogrammatical defectiveness in units, non-compositionality, i.e., the meaning arising from word-by-word interpretations differs from the accepted meaning of the unit, and institutionalized status, i.e., being recognized and accepted as a lexical item of the language (Moon 2003: 6–8).

Toponymic idioms constitute a group of units containing culture-specific components; therefore, such phrases are particularly interesting in a contrastive perspective, in which two or more languages are compared. It is assumed that:

1. Idiomatic expressions containing toponyms, components functioning as culture carriers, tend to be axiologically loaded.
2. When such toponymic idioms are analyzed in a contrastive perspective, the question arises how axiology is expressed by toponymic idioms of various languages and whether the similarities or differences dominate the comparative picture of the axiology conveyed by the idioms of particular languages.

1.1 The cultural character of idioms

Idioms of a given language tend to reflect the culture, illustrating the correlation between language and culture (Teliya et al. 2001: 55). Therefore, to provide an in-depth analysis of idiomatic units, interdisciplinary studies, especially ethnolinguistic analyses allowing for exhaustive treatment of the issue, have to be conducted in order to ensure a proper approach to idiomatic expressions viewed as carriers of cultural connotations.

The relations between language and culture are complex, since the language expresses, embodies and symbolizes cultural reality (Kramsch 2000: 3); thus, idioms can be analyzed from various points of view. For instance, it is possible to

distinguish certain groups of components, which tend to be cultural-bound elements in the majority of languages, both European and non-European.

Such culture-specific components are, for instance, the names of material realia (clothes, musical instruments, dishes, e.g., *bigos* 'the Polish national dish, a stew made from alternate layers of sauerkraut and meat', cf. Ayto 2002: 29), the names of social realia (*kolhoz* 'a Soviet collective farm'), faunal and floral terms particular to the area occupied by a given ethnic community (e.g., *kangaroo* in the Australian variety of the English language; cf. Fernando 1996: 92–93), and proper names (*Campidoglio*, the place associated with the Italian government; Castoldi and Salvi 2003: 68).

The last group mentioned above contains various kinds of nomina propria, i.e., proper nouns; for instance, anthroponyms (names of persons, e.g., *John*), toponyms (names of places, e.g., *London*), ideonyms (names of books, films, etc., e.g., *The Newsweek*) and zoonyms (names of animals and pets, e.g., *Fido*).

1.2 The toponym as an axiologically marked idiom component

After anthroponymic idioms, toponymic idioms, i.e., the units containing a place name either natural or social (cf. McArthur 1996: 704), tend to constitute a very important group of idioms in most European languages (Spagińska-Pruszak 2003: 77). In fact, a great number of proper names, toponyms included, as well as a number of other groups of nouns, for instance, faunal terms (Szerszunowicz 2005), function as evaluatively loaded units.

It is assumed that the core component, i.e., the toponym, used figuratively, evokes concrete connotations in native language users, either carrying the axiological load itself or indirectly, by means of the combined meaning of the whole unit. In the former case, the presence of toponymic components in axiological idioms of a given language renders it possible to present a 'map' reflecting the axiology contained in the phraseology, since certain toponymic components have positive connotations in the collective memory, while others are depreciative. In the latter case, the axiological toponymic idioms tend to reflect extralinguistic factors of universal, national or local character, which is also important in terms of the translation of such units. The evaluative power may also be based on the exploitation of secondary associations of the toponym, its phonetics and other factors of lesser importance.

Furthermore, according to Chlebda (2003: 245–254), the majority of idiom occurrences excerpted from a variety of texts, both written and oral, show that the units at issue tend to appear as modifications, differing from their base, canonical forms. Therefore, as to toponymic idioms, it can be assumed that the language

users' knowledge of the connotations of toponymic components enhances the modification potential of phraseological units containing place names.

1.3 The contrastive analysis of the axiologically marked toponymic idioms: objectives

Certain toponymic idioms reflect the evaluation of people, places, phenomena, behaviour, etc. present in the collective memory of a given nation. Presenting a map of axiologically loaded toponymic components of idioms renders it possible to analyze which names tend to be carriers of evaluation in the language analyzed as well as indicating which factors contributed to their acquiring a symbolical value. In a contrastive perspective it is of great interest to what extent such axiological maps overlap; furthermore, how the differences between the maps of two languages influence the translation equivalence of such idioms. Diachronic studies regarding the issue of the axiology of toponymic idioms may show interesting differences of significant changes in the maps from different periods, whereas synchronic research is particularly important from the point of view of translation studies, lexicography and language pedagogy.

The material used for the analysis is composed of Polish and Italian units excerpted from the dictionaries, axiologically marked idioms, whose components are the toponyms. The Polish corpus consists of 37 units, while the Italian one is composed of as many as 98 idioms. The two languages were chosen, since bilingual dictionaries of idioms, (Drzymała 1993; Mazanek, Wójtowicz 1993; Salwa, Śleszyńska 1993; Zardo 2002; Podracka 2006), contain very few idioms from the vast majority included in the contemporary monolingual Polish and Italian dictionaries of idioms, from which idioms have been excerpted. Therefore, the translation of units carrying axiological load, which are not included in bilingual dictionaries, is particularly difficult since it requires from the translator knowledge about the axiological potential of a given unit. In the case of idioms realizing similar patterns the possibility of mistranslation exists (cf. Szerszunowicz 2006a). A proper lexicographic description should provide sufficient information of the axiological character of the unit.

Apart from Italian and Polish, in some cases idioms from other languages are presented as well to exemplify issues in a more detailed way. The term *axiological markedness* is used in the broad sense of the word, since all toponymic idioms performing an evaluative function are included. The units containing toponymic derivatives are excluded from the corpora. The objective of the analysis is to discuss how evaluation is expressed in the two languages compared as well as to find some models of the relation between the L1 idiom and its equivalent in L2.

2. The evaluative effect of toponymic idioms

The evaluative effect of toponymic idioms can be created by a number of mechanisms. For instance, the toponym can be a core evaluative means. The toponymic component may also be reinterpreted etymologically, which results in the idiom evaluative load. Evaluation can be expressed by the combined meaning of idiom components. Moreover, there are some means of intensifying the evaluation expressed by toponymic idioms.

2.1 The toponym as the core evaluative means

The analysis of axiologically loaded toponymic idioms shows that evaluation is expressed by means of various mechanisms. The first is expressing the valuation by referring to the symbolical value of the toponymic component in a given culture; in other words, a given evaluation is attributed to the toponym. The evaluation is automatically associated with the name by language users, thus becoming a conventional metaphor. In fact, in a number of toponymic idioms the core evaluative element is the toponym itself, since the name possesses a symbolic meaning which functions in the collective memory of a nation or a local community.

Therefore, the toponyms evoke definite connotations even if not used as components of an idiom; for instance, in many languages the names of places where a mental hospital is situated carries a negative valutation. The Polish names of cities and towns where such a hospital is located, such as *Choroszcz, Drewnica, Jarosław, Świeć, Tworki* and many others express a negative evaluation, connoting stupidity, abnormality, craziness, etc. The toponyms listed appear in the realizations of the model 'go to + the name of the town/city where a mental hospital is', likewise the French name *Chartenton* or the Italian toponym *Aversa*.

Similarly, in most languages there are funny-sounding place names, both real and fictitious, which symbolize a remote, old-fashioned, retarded place, for instance, the Polish name *Pipidówka* or the Italian one *Carrapipi*. Such toponyms also appear in the realizations of the schema 'X from Y', where X is the person/thing and Y is the place name conveying a negative evaluation. For example, the Polish phrase *Anglik z Kołomyi* (lit. an Englishman from Kołomyja, 'a person pretending to be an English man; a person trying to apper more important') and the Italian idiom *casaliga di Voghera* (lit. housewife from Voghera, 'a narrow-minded woman') realize the schema at issue. In both cases, the place name is the carrier of negative evaluation since the names are synonyms of a small, godforsaken place. Such idioms, which are realizations of the same schema, but differ in meaning, should be treated as potential false friends (cf. Szerszunowicz 2006a).

Positive evaluative load is in turn expressed by the names of the places from which the authorities govern the country. For example, the Italian name *Campidoglio*, not only in Italian, but also in many other languages, especially in journalese, is a synonym of the place where the political power of the nation is based (Castoldi and Salvi, 2003: 68). Vertical, up-and-down evaluation is observed in idioms such as *dalla Rupe Tarpea al Campidoglio* or the American English *from the log cabin to White House*. Such components as *Rupe Tarpea, Campidoglio* or *White House* are the key words in the analysis of the axiologically loaded phraseological units in the collective memory of a given nation.

2.2 The reinterpretation of the toponymic component

Furthermore, in the case of toponymic idioms, the evaluation can be expressed by the etymologic reinterpretation of the onomastic components (Randaccio 2006; Szerszunowicz 2006b). Certain allusive place names have a playful reference characterizing a place by means of exploiting their derivational structure and meaning, thus introducing an element of word play based on the structure of a given name (McArthur 1996: 32; cf. Baldick 1991: 6). For instance, in the case of the Italian idiom *cavallo dell'Asinara* (lit. *the Donkeyville horse*), evaluation is expressed by the etymologic reinterpretation of the onomastic components since the derivational base of the place name, i.e., asino 'donkey', has a conventional pejorative metaphorical meaning. The *–ara* suffix is common in the Italian toponymy (cf. Queirazza et al. 1997). However, the majority of the reinterpreted allusive names, whose ornamental character is actualized in a given unit, are authentic, with relatively fewer fictitious names.

The idioms of jocular nature constitute a particularly interesting group of axiologically loaded fossilized units, since they express the negative evaluation in a humorous way, thus changing the style. For instance, in the Italian language numerous idioms are used to evaluate products as of low quality, e.g., steel (*essere acciaio di Ferrara*, *ferro* 'iron',), diamonds (*diamanti di Vetralla*, *vetro* 'glass'), butter (*burro di Segovia*, *sega* 'suet'), fruit (*frutta di Marcialla*, *marcio* 'rotten'), cigarettes (*cigarette di Cartagine*, *carta* 'card'). However, the depreciative character of the idioms is ameliorated with humor; therefore, the negative evaluation is of different shade from the markedness of other groups.

2.3 The combined meaning of idiom components

Apart from the direct axiological markedness of toponyms, in some cases the evaluation is expressed by the summaric meaning of all components constituting the idiom. The toponym itself does not possess the metaphorical meaning and tends

not to be used as an allusive name on its own. For instance, the Italian idiom *portare frasconi to Vallombrosa* (lit. bring branches to Vallombrosa) expresses an evaluation of an activity, describing it as a pointless, and the negative evaluation is created by means of combining elements. In the Italian language the Latin model *in sylvam ligna ferre* (lit. bring wood to the forest) is realized with reference to the common knowledge that Vallombrosa, situated in Tuscany, is rich in forests. Therefore, the axiology is expressed as a sum of elements constituting the meaning of the idiom: 'carry branches to Vallombrosa', whose onomastic component is the name *Vallombrosa* – the place known for its forests, i.e., to be unreasonable, stupid, pointless, etc. The name *Vallombrosa* is not used as a conventional metaphor on its own.

Furthermore, it is observed that in some cases the appearance of a toponym in an idiom of evaluative character is based on the rhyme requirements. In some cases the phonetics of the toponym may be one of the means of expressing the evaluation. For instance, in the Polish idiomatic phrases *Francja elegancja* (lit. France elegance), the toponym France is associated with the capital of the fashion world, and it rhymes with the noun *elegancja*. In other idioms phonetic value is the dominant feature of the evaluative potential realized in the unit; for example the Polish unit *życie jak w Madrycie* (lit. life as if in Madrid) contains the component *Madryt*, which does not evoke any particular stereotypical image in the language users.

It is worth noting that one of the Polish idioms describing the appearance *Jedno oko na Maroko, a drugie na Kaukaz* (lit. one eye at Marocco, the other at the Caucasus 'with a very bad squint'), the components *Morocco* or *Caucasus*, the latter one sometimes substituted with the name of a Polish city, *Zgierz*, do not carry any evaluative value themselves, yet the whole idiom gains an evaluative character, since there is added meaning in comparison with the neutral equivalent *zezowaty* 'cross-eyed'. It should be stressed that in some dictionaries the onomastic components are treated as appellative nouns, which is reflected in the orthography of the place names not written with initial capital letters (*Jedno oko na maroko, a drugie na kaukaz*).

2.4 Means of intensifying the evaluation expressed by toponymic idioms

In some cases the evaluative power of toponymic idioms may be increased by a variety of means. First of all, the language user can exploit linguistic means to enhance the axiological load of the idiomatic expresion. One of the commonest means is the insertion of intensifying words; for instance, *coś jest czyimś Waterloo* (lit. something is somebody's Waterloo), *coś jest czyimś prawdziwym Watreloo* (lit. something is somebody's real Waterloo). Axiologically marked adjectives and adverbs, for instance 'terrible', 'horribly' etc., contribute to the evaluative power of a given unit. Neutral words or word combinations can also be added in order

to increase the evaluative load of the unit; for example, *jechać do Choroszczy* (lit. go to Choroszcz), *jechać do Choroszczy na sygnale* (lit. go to Choroszcz with the siren on). It is also possible for language users to insert positively marked words to add irony; for example, *cavallo dell'Asinara* (lit. Donkeysville horse), *il bel cavallo dell'Asinara* (lit. a beautiful Donkeysville horse). Moreover, certain components can be substituted or omitted, which in some cases may influence strongly the evaluative power of the units at issue.

Moreover, it is worth observing that extralinguistic means may strongly influence the evaluative force of a unit. The implementation of gesture, together with an appropriate intonation and accentuation, plays an important role in expressing the axiology by means of an idiom. In the case of a few idioms particular gestures are commonly associated with the phrases; yet, the use of the gesture is only an optional means of increasing the evaluative force. For instance, in the Italian language there are numerous idioms expressing a negative evaluation of a betrayed husband, in which the allusive toponymic components exploit the word play on the noun *corn* ('horn'). Such idioms, for example, *mandare uno a Cornetto* (lit. send somebody to Hornville), *mandare uno in Cornovaglia* (lit. send somebody to Cornwall), *andare to Cervia* (lit. go to Deerville), based on an onomastic pun (Lurati 2002: 23), tend to be accompanied with a particlular gesture, i.e., showing horns with the index finger and little finger (Diadori 1990: 55).

3. The cross-linguistic equivalence of axiologically marked idioms

From the point of view of the cross-linguistic equivalence of axiologically marked idioms, the following groups of equivalents are distinguished: absolute equivalents of axiologically marked idioms, equivalents of axiologically marked idioms with substituted toponymic components, equivalents of axiologically marked idioms without toponymic components, equivalents of axiologically marked idioms with recreated toponymic components, non-idiomatic equivalents of axiologically marked idioms.

3.1 Absolute equivalence of axiologically marked idioms

From a contrastive perspective, axiologically marked toponymic idioms can be classified according to the degree of their cross-linguistic equivalence. The units with the highest degree of equivalence are composed of toponyms whose connotations are universal; i.e., they originate from the common background, are connected with events of particular significance (e.g., *Waterloo*), and are universally known to the vast majority of language users. The name gains a symbolical value, used in numerous idioms and on its own.

Therefore, there are idioms having close equivalents in L2 (identical structure, meaning, toponymic component; e.g., following the Latin pattern *mari aquam addere* (lit. bring water to the sea): English: *bring owls to Athens*, German: *Eulen nach Athen tragen* (lit. bring owls to Athens), Italian: *portar nottole ad Atene* (lit. bring owls to Athens), Polish: *nosić sowy do Aten* (lit. bring owls to Athens). It is worth emphasizing that such units refer to the common European linguistic and cultural heritage. Many of them tend to be used infrequently (cf. Moon 1998: 41). The majority of idioms constituting this group tend to be archaic or formal.

Moreover, it is worth emphasizing that in certain cases the idioms may be recreated as literal translations in the target language and they will convey identical evaluations in L1 and L2, since the evaluative markedness of the toponymic components is of universal character. For instance, the idiom included in Polish dictionaries of phraseological units, *coś jest czyimś Waterloo* (lit. something is somebody's Waterloo; labled *formal*) is translatable into the Italian language, as well as many other languages, without the loss of the axiological element. Taking into consideration that the idiom is labeled *formal*, it can be assumed that in particular contexts such a translation may be perfectly acceptable and functional.

3.2 Equivalents of axiologically marked idioms with substituted toponymic components

The next group is composed of idioms whose equivalents realize the same structural and semantic model while the toponym is replaced with a toponymic counterpart of national character. Thus, instead of universal reference, there is reference to shared knowledge within a group. For instance, the Latin model *mari aquam addere* is realized in a number of languages following the pattern 'carry water to + the name of a long river', with the implication 'there is a lot of water in the river, so it is pointless to do so', e.g., English: *carry water to the Thames*, Romanian: *cara apa-n Dunare* (lit. carry water to the Danube), Polish *do Wisły wodę nosić* (lit. carry water to the Vistula), German: *Wasser in den Rhein tragen* (lit. carry water to the Rhein). Moreover, the toponym of national character may also belong to a class different from hydronyms; for instance, it can be a city name, as it is in Russian: *yekhat' v Tulu so svoim samovarom* (lit. go to Tula with one's samovar), English: *bring coals to Newcastle*, Spanish: *ir con naranjas a Valencia* (lit. go to Valencia with one's oranges).

It is worth emphasizing that apart from the realizations of the above types with universal and national components, local colour can be introduced by the substitution of the universal/national element with a local one; for instance, in Italian: *portar acqua ad Arno* (lit. carry water to the Arno), idiom used in Tuscany. Another example is 'go to + the name of the town/city where a mental hospital is

situated', whose realization are found, among others, in Polish, German, Italian, Croatian and Czech using local toponyms.

Local reference is usually unknown to the inhabitants of other regions: decoding of the axiology expressed by the idiom thus requires shared knowledge. The units realizing a model with the substituted local color component are equivalent within one language, forming a synonymous chain of realizations in one language in which the onomastic component varies depending on the region. It should be mentioned that such idioms have various statuses: some of them are included in lexicographic works and appear in the standard variety of the language, while the use of others is limited to a small group of inhabitants of a particular region.

3.3 Equivalents of axiologically marked idioms without toponymic components

The equivalents of some axiologically loaded idioms are units of idiomatic character, expressing an evaluation, but they do not contain a toponym. For instance, the Italian idiom *fare come l'angelo di Badia* (lit. behave like the angel of Badia, i.e., the angel on the dome of Badia, in Florence, which keeps changing the position according to the direction of the wind, 'keep changing opinions') possesses two idiomatic equivalents in the Polish language, both in the form of similes.

The Polish equivalents are based on the same model, and, in fact, their pictorial character is similar since the angel is substituted with the flag and the rooster (*być jak chorągiewka na wietrze* lit. be like a little flag in the wind; *być jak kurek na kościele* lit. be like a little rooster on the church roof); yet the images presented by the Polish idioms are very general, whereas the Italian one conveys a concrete, specific reference to the local feature of the Florence area. Moreover, one should take into consideration the fact that the unit *być jak kurek na kościele* is an idiom of low frequency in the contemporary Polish. Thus, it can be concluded that the evaluative power of the Italian idiom is greater than the axiological force of the Polish units.

3.4 Equivalents of axiologically marked idioms with recreated toponymic components

Another group is formed by idioms which can be re-created in the target language as non-conventional phrases of similar stylistic markedness, since the toponymic element is an etymologically reinterpreted element of axiological character (Italian: *cavallo dell'Asinara*, English: *the Donkeysville horse*). Idioms whose components are etymologically marked occur with a greater frequency in Italian than in Polish; in fact, in this group absolute equivalents are not observed.

It should be stressed that in the case of such toponymic idioms the way axiological value is expressed relies heavily on the word play, which is retained if the allusive toponym is re-created. Thus, the device of expressing axiology is identical in terms of the structure. What is lost in the process of translation is the conventional character of the unit since in the target language the phrase with the re-created toponym is innovative.

3.5 Non-idiomatic equivalents of axiologically marked idioms

The last group of axiologically loaded toponymic idioms is composed of idioms whose equivalents are non-idiomatic and do not contain any toponymic components. Equivalence can be illustrated with the Italian idiom *portare soccorso di Pisa* (lit. bring Pisa help) and its Polish equivalent *przybyć z pomocą zbyt późno* (lit. come with the help too late). As it can be seen from the comparison of this pair of idioms, the meaning of the Polish word combination is identical in terms of the general meaning, yet the original unit and the one in the target language differ significantly in their style. The absence of the toponymic component eliminates the cultural character of the unit since the target language item is devoid of any cultural reference.

The toponymic component evokes certain associations in the language user, to a greater or lesser extent, depending on the degree of lexicalization. The idiom containing the toponymic component is cultural, having universal, national or local symbolic value, while the non-idiomatic realization does not. Moreover, it loses the conventional, fossilized character the idiom has, which, in consequence, results in diminished axiological value expressed by the non-idiomatic phrase in the target language.

4. Conclusions

The analysis of the axiological toponymic idioms excerpted from lexicographic works shows that numerous idioms function as carriers of evaluation. The toponym can be the core element of evaluation if it possesses a symbolical value in a given culture. Such toponymic components can be roughly classified into three groups: universal, national and local. Yet, it should be stressed that the onomastic component at issue does not have axiological import in all cases. It is observed that toponyms lacking in symbolic value can also gain one within idiomatic expression, for instance due to etymological reinterpretation.

As to the translatability of the axiologically marked idioms, the following groups can be distinguished: units possessing absolute equivalents (idioms

realizing one model, in which the toponym remains the same in the target language, since the axiological markedness is universal); the units with close and fairly close equivalents (idioms realizing one model with substitutions, in which the toponym can belong to the same class as in the L1 or may be different, thus changing slightly the pictorial character of the unit); units re-created in L2, thus losing their conventional character (mostly those with allusive names based on etymological reinterpretation and in some cases with national or local components); units possessing idiomatic equivalents without the toponym; and units possessing non-idiomatic, descriptive equivalents in L2.

It is worth observing that in the languages compared few idioms containing toponyms are devoid of any evaluative markedness, since the vast majority express bipolar axiology, i.e., differentiating between the good and the bad, with prevalence of negative evaluations. The vast majority of L1 idioms can be expressed in L2 by means of an idiomatic structure with a different toponymic component or an idiom which does not contain a place name.

The fact that very few Polish idioms possess absolute equivalents in Italian and vice versa shows that the axiological markedness of idioms undergoes certain changes in the process of translation. Idioms of humorous character may retain their axiological effect if re-created in L2. In terms of equivalence, the greatest loss in axiological character of an idiom is observed in units whose equivalents are non-idiomatic, since a descriptive equivalent has much less evaluative power. The final conclusion is that the analysis of the axiological markedness of particular groups of idioms contributes to the lexicographic description of units as well as to the proper translation of such expressions.

References

Ayto, John. 2002. *An A-Z of food and drink*. Oxford: OUP.

Baldick, Charles. 1991. *The concise Oxford dictionary of literary terms*. Oxford: OUP.

Castoldi, Massimo & Ugo Salvi. 2003. *Parole per ricordare. Dizionario della memoria collettiva*. Bologna: Zanichelli editore.

Chlebda, Wojciech. 2003. *Elementy frazematyki. Wprowadzenie do frazeologii nadawcy*. Łask: Oficyna Wydawnicza LEKSEM.

Diadori, Pia. 1990. *Senza parole. 100 gesti degli italiani*. Roma: Bonacci editore.

Drzymała, Piotr. 1993. *Fraseologia italiana. Słowniczek frazeologiczny włosko-polski*. Poznań: Wydawnictwo Naukowe UAM.

Fernando, Chitra. 1996. *Idioms and idiomacity*. Oxford: OUP.

Kramsch, Claire. 2000. *Language and culture*. Oxford: OUP.

Lurati, Ottavio. 2002. *Per modo di dire … Storia della lingua e antropologia nelle locuzioni italiane ed europee*. Bologna: CLUEB.

Mazanek, Anna & Janina Wójtowicz. 1993. *Idiomy polsko-włoskie. Fraseologia idiomatica polacco-italiana.* Warszawa: Wydawnictwo Naukowe PWN.

McArthur, Thomas (Ed.), 1996. *The Oxford companion to the English language*. Oxford: OUP.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English.* Oxford: OUP.

Podracka, Maria K. 2006. *Idiomy włoskie*. Warszawa: Wydawnictwo REA.

Quartu, Monica B. 2000. *Dizionario dei modi di dire della lingua italiana.* Milano: RCS Libri.

Queirazza Giuliano G. et al. 1997. *Dizionario di toponomastica. Storia e significato dei nomi geografici italiani.* Torino: Utet.

Randaccio, Roberto. 2006. Toponomastica allusiva: luoghi reali e fantastici nelle locuzioni evocative (nei detti proverbiali, nei lessici e in letteratura). In *Lessicografia e onomastica. Atti delle Giornate internazionali di Studio Università degli Studi Roma Tre, 16–17 febbraio 2006*, P. D'Achille & E. Cafarelli (Eds), 147–158. Roma: Società editrice romana.

Salwa, Piotr & Małgorzata Śleszyńska. 1993. *Wybór idiomów włoskich.* Warszawa: Wiedza Powszechna.

Spagińska-Pruszak, Agnieszka. 2003. *Intelekt we frazeologii polskiej, rosyjskiej i chorwackiej*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Szerszunowicz, Joanna. 2005. The axiology of faunal metaphors in Italian and Polish. *Jaunųjų Mokslininkų Darbai. Young Researchers' Works* 3(7): 196–200.

Szerszunowicz, Joanna. 2006a. Pseudo-equivalents in English, Italian and Polish faunal phraseology. In *Proceedings XII EURALEX International Congress. Atti del XII Congresso Internazionale di Lessicografia, Torino 6–9 settembre 2006*, Vol. 2, E. Corino, C. Marello & C. Onesti (Eds), 1055–1060. Alessandria: Edizioni dell'Orso.

Szerszunowicz, Joanna. 2006b. On allusive place names in Italian idioms. *Jaunųjų Mokslininkų Darbai. Young Researchers' Works* 1(8): 179–182.

Teliya, Veronika et al. 2001. Phraseology as a language of culture: Its role in the representation of a collective memory. In *Phraseology. Theory, analysis, and applications*, A.P. Cowie (Ed.), 55–75. Oxford: OUP.

Zardo, Manuela. 2002. *Langenscheidt 1 000 idiomów włoskich*. Warszawa: Langenscheidt Polska.

PART III

Historical change

The role of prefabs in grammaticization

How the particular and the general interact in language change

Joan Bybee & Rena Torres Cacoullos
University of New Mexico

Abstract

Studies of grammaticization often reveal skewed distributions of lexical items in grammaticizing constructions, suggesting the presence of prefabs using these constructions. We examine here the role of prefabs in the grammaticization of *can* in English and the progressive *estar* 'be (located)' + V-*ndo* (Gerund) in Spanish. The data suggest that prefabs play a role in advancing formal and semantic change. We argue that (1) prefabs are ahead of the general construction in unit-hood status in early stages and thus demote the independent lexical status of the emerging grams, and (2) in their association with semantic classes of which they are the most frequent member, prefabs promote the productivity of the general construction. The evidence shows that prefabs maintain associations with the related general construction.

1. Introduction

The fact that natural discourse relies heavily on repeated, conventionalized multiword strings has implications for processing theories, production theories and grammatical theory. In this paper we explore the interaction of specific conventionalized multi-word strings, which we will call "prefabs" (following Erman and Warren 2000), and the more general constructions that make up the grammar of a language. In particular we focus on the way prefabs participate in the process of grammaticization by studying prefabs that have developed along with the Progressive construction with *estar* in Spanish and the auxiliary construction with *can* in English.

Our theoretical perspective is that of usage-based construction grammar in which cognitive representations are affected by the speaker's experience with language (Goldberg 2006; Bybee 2006). Tokens of experience are represented in memory as exemplars of varying strengths. The representations of constructions consist of categories that group together all the exemplars of a given construction, based on semantic and formal similarity.

The model of lexical associations proposed for morphologically complex words in Bybee (1985, 1988, 2001) can be extended to multiword units and constructions. In this model, associations made among related forms are gradient and depend upon the degree of semantic and phonological similarity and the token frequency of the specific items (as we explain below). One of the main determinants of memory storage is frequency in experience; thus specific instances of constructions may occur as units in memory storage, even if their meaning and form is predictable from the more general construction. An expression such as *that drives me crazy* may occur as a unit of storage and may be accessed in one step. However, when stored units are themselves complex they can still be related in representation to the smaller units that comprise them (*that, drives, me,* and *crazy*) as well as to the general construction the stored expressions instantiate.

In this view, there is no discrete division between fixed expressions and productive formations, rather, these two types of linguistic expressions form the two poles of a continuum. Evidence of the continuum between the processing of fixed expressions and productive constructions includes the fact that even highly fixed expressions sometimes undergo expansion, as when a radio news reporter was heard to say *all chaos broke loose*. One might have thought that *all hell broke loose* was entirely fixed, but he was able to make a substitution inside this expression. Some expressions allow for considerable expansion, as when the adjectives that follow *drives someone* + ADJ are studied in a corpus. It is found that *mad, crazy, insane, wild, nuts, up the wall* and several others are possible in this construction (Boas 2003).

Moreover, constructions that are often thought of as rather general and schematic nonetheless often have lexical restrictions, as the ditransitive construction occurs only with a certain set of verbs, such as those denoting 'giving': *give, pass, had, sell, trade, lend, serve, feed* and other classes as well (Goldberg 1995: 126).

Given this continuum, we can identify the dimensions that determine the variation from one pole to the other. In this paper we present evidence for the following three dimensions.

i. Productivity: If the expression is schematic, that is, if it has slots that can be filled by a class of items, then it will also vary on a scale of productivity depending upon the number of types that can occupy its open position and the semantic generality of the class.
ii. Transparency of meaning: fixed expressions can have fully compositional meaning, as in expressions such as *open the door* or *pass the salt*. Less than transparent meaning occurs in idioms that have a metaphoric interpretation (e.g., *pull strings*) or in frequent expressions that have developed some pragmatic or semantic nuances or changes that distance them from the more general constructions with which they are related (e.g., *I don't know* as a discourse marker [Scheibman 2000]).
iii. Analyzability: expressions may also differ in the extent to which the units composing the expression are associated with the etymologically same units in other constructions. Diagnostics for analyzability include the ability to add modifiers or other elements that separate the units of the expression or to appear in different constructions, as for example, when the elements are recomposed into a passive.[1]

Many researchers propose two modes of processing to underlie the Open Choice Principle and the Idiom Principle (as Sinclair 1991 put it; see Erman and Warren 2000; Van Lancker [this volume]; Jackendoff 2002), despite the gradient between monomorphemic units on the one hand and conventionalized, multiword sequences on the other that we have just described and for which we present further evidence below. This gradience suggests that two distinct types of processing are not involved. In contrast, we propose that the access of stored units in production and perception is the same process whether the units are simple or complex; the observed gradience is not a property of the type of processing but

1. See Langacker 1987 for a discussion of analyzability (pp. 292–298; 457–460) and compositionality or transparency of meaning (452–457).

rather of the length, complexity and degree of fixedness of the stored units. Thus accessing the stored linguistic representations is essentially the same whether the unit is a monomorphemic word, such as *wall*, a phrase such as *the wall* or a partially schematic construction such as *X drives me Y*, where the X position can be filled with almost any sort of NP and the Y position contains an adjective or prepositional phrase from the class related to *mad*, *crazy*, *up the wall*.

Besides degrees of complexity in storage, another source of complexity in utterances arises from the fact that the schematic slots in constructions can themselves be filled with either simple or complex material. Dąbrowska and Lieven (2005) use the term "superimposition" to describe the process by which an accessed unit is used to fill a position in a partially schematic accessed unit or construction. To use Dąbrowska and Lieven's example, *shall I* PROCESS*?* (where PROCESS stands for the set of verbs or verbal complexes that may occur in that position) and *open that* can be superimposed to derive the expression *shall I open that?* All the properties of the two units – their phonetic form, meaning and pragmatics as derived from previous experience are carried along in the process of superimposition. Note that units involved in superimposition may in themselves be the result of superimposition, as in the example *open that* which was derived by superimposing *that* and *open* OBJECT. Thus the process of superimposition is one of the sources of syntactic complexity in utterances; the other source is the complexity that is inherent to the stored unit.

Given this general framework coupled with an exemplar model of linguistic representation, usage data suggests that certain exemplars of constructions have differential representation depending upon their frequency of use (Bybee 2003, 2006). One of our interests in this paper is to examine how specific exemplars of constructions affect the overall meaning and use of the construction. We cast this question in a diachronic context and examine the way conventionalized instances of constructions or prefabs interact with the more general construction as grammaticization proceeds. Rather than viewing prefabs as something distinct from and perhaps peripheral to grammar in the traditional sense, we argue that prefabs constitute important loci of grammatical development in the diachronic domain. By implication, such conventionalized expressions have important interactions with more general constructions in the synchronic domain.

2. Cognitive consequences of skewed frequency distributions in constructions

Corpus-based studies of constructions reveal an uneven topology for the distribution of lexical items in constructions. In many cases, one or a small number of

lexical items occur frequently in the construction and other lexical items occur once or twice in the construction. Thus Goldberg, Casenhiser & Sethuraman (2004) find that in mothers' speech to children aged 20 to 28 months certain verbs occurred frequently in certain constructions: for instance, *go* accounted for 39% of verbs in SUBJ VERB OBLIQUE constructions; *put* accounted for 38% of all SUBJ VERB OBJECT OBLIQUE constructions and *give* 20% of all SUBJ VERB OBJ OBJ_2 constructions.

Bybee & Eddington (2006) studied Spanish change-of-state verbs and the adjectives that accompanied them and found that certain pairings were of very high frequency, e.g., *quedarse solo* 'to end up alone'; *quedarse quieto* 'to become still'; *quedarse sorprendido* 'to be surprised'; *ponerse nervioso* 'to get nervous'. These expressions are prefabs in that they represent the normal, conventionalized way of expressing certain commonly-referred to changes of state. It was also found that these expressions formed the centers of exemplar categories, as the corpus also contained many single examples that were related semantically to these more frequent expressions. Thus the prefabs play a central role in determining the range of use of the constructions. See Wilson (this volume) for details about the diachronic development of these constructions from prefabs.

Goldberg and colleagues (Goldberg, Casenhiser & Sethuraman 2004, 2005; Casenhiser & Goldberg 2005) argue that the skewed distribution in constructions aids in acquisition because the frequent expressions or prefabs play a crucial role in helping the child grasp the meaning of the constructions. They designed an experiment to test the contribution of type and token frequency in which both children and adults were taught a nonce argument structure construction in English. The construction had a nonce verb (with a suffix in some of the conditions) and the verb appeared at the end of the clause. The meaning of the construction was taught through a video presentation that accompanied the linguistic stimuli. In one condition nonce verbs appeared in the stimuli with the same token frequency, while in the other condition the same number of verbs was presented, but one had a higher token frequency than all the others. In the latter condition, learning was more successful. The hypothesis about the facilitation of learning is that the repetition of a particular verb in a particular construction helps to establish the correlation between the meaning of the construction and its formal expression. Goldberg (2006) goes on to demonstrate that in category learning in general a centered, or low variance, category is easier to learn. The condition with one instance of higher token frequency is just such a category.

Lieven and colleagues (Lieven et al. 1997; Lieven et al. 2003; Dąbrowska & Lieven 2005; Lieven et al. this volume) demonstrate that early children's utterances are strongly based on utterances the children have experienced before, in

their own speech or in the speech of adults. Dąbrowska & Lieven (2005) argue that children start their acquisition of grammar with multiword sequences that are rather fixed and repeated verbatim and gradually learn to substitute lexical items into the slots in the construction represented by the sequence. Thus the analysis of the repeated utterances and the build-up of more abstract and schematic constructions emerges gradually out of experienced and repeated tokens. However, even after the more abstract constructions are established (say, in adults), many utterances may still be produced by accessing large, pre-assembled and lexically-specific sequences from memory.

These studies, then, all show a significant interaction of prefabs with more combinatorial tokens of constructions. This is possible because prefabs have not necessarily lost their internal structure, nor have their component parts necessarily lost their identities. Nunberg et al. (1994) argue that many phrases taken as idioms actually retain their compositionality in the sense that their parts "carry identifiable parts of their idiomatic meanings" (496). In addition, such "idiomatically combining expressions" retain their morphosyntactic analyzability. Thus it is argued that even in expressions with unpredictable meaning, such as *pull strings*, the two words each still contribute to the idiomatic meaning in the sense that one can identify for any given case what or who were the "strings" and what was done to "pull" them. So if such idioms have discrete parts that are associated with other VERB – OBJECT constructions as well as with other instances of the lexical items involved, then other sorts of prefabs can certainly have these properties as well. That is, despite holistic processing and chunk-like storage, prefabs can still be related (to varying degrees) to the words and constructions of which they are constituted. It follows then, that in language change, prefabs might have an impact on the nature and rate of change in constructions.

It is known from studies of discourse variation and grammaticization that increasing token frequency of an expression leads to increasing opacity of internal structure and increasing autonomy from the more general construction, which enables the resulting single processing unit to gain new discourse-pragmatic functions (Bybee 2003: 618; cf. Thompson & Mulac 1991; Company 2006; Torres Cacoullos 2006). Nevertheless, we argue that prefabs can maintain associations of gradient strength with the more general construction unless and until increases in frequency and concomitant semantic/pragmatic change reach high levels.

In a study of current variation reflecting ongoing grammaticization, Torres Cacoullos & Walker (2009) showed that the patterns affecting the general construction also affect fixed formulas: even though prefabs develop their own discourse-pragmatic characteristics, they retain traces of the constraints on their associated construction. These researchers used multivariate analysis to discover

a number of language-internal factors conditioning the variation between *that* presence and absence in naturalistic speech data. *I think, I guess* and a handful of other frequent 1st person singular and Present tense collocations (*I remember*, *I find*, *I'm sure*, *I wish*, *I hope*) have become conventionalized as discourse formulas that function more as epistemic or evidential adverbial phrases than as main-clause propositions (e.g., Thompson & Mulac 1991; Diessel & Tomasello 2001; Thompson 2002). Torres Cacoullos and Walker found that even though the rate of *that* with prefabs *I think*, *I guess* is low, the linguistic conditioning parallels instances of the more general construction with more robust variation: the two strongest constraints, intervening material and type of subject, are both operative, and with the same direction of effect (the presence of intervening material and full NP subjects favor *that* presence). They argue that not only do grammaticizing constructions retain lexical meaning (Bybee & Pagliuca 1987; Hopper 1991), but prefabs retain grammatical properties, manifested in the parallelism of constraints on variation. Thus, the units of formulaic language maintain associations with productive constructions, contra the view that would isolate the former in a lexicon separate from the grammar.

Other studies of grammaticization have also revealed skewed distributions of lexical items in grammaticizing constructions. It is often noted that grammaticization gets its start in constructions with particular classes of items. For instance, Carey (1994) finds that the Old English resultative construction that becomes the Perfect was used most frequently with verbs of mental state and reporting verbs and its meaning first conventionalized in expressions with these verbs. Thus to study both the meaning changes in grammaticization and the way grammaticizing constructions expand and generalize, it is instructive to examine the use of such constructions in prefabs.[2]

If prefabs are processed more holistically than more compositional word combinations, the meaning of the individual units making up the expression will be less transparent. We regard the effect of holistic processing to be cumulative; the more often a sequence is accessed as a whole unit, the stronger the path to that type of access will become (Hay 2001). We will argue that the cumulative effect of this more holistic processing contributes to the pragmatic and semantic changes that occur in grammaticization. Our consideration of diachronic data on the development of the English auxiliary *can* from the Old English verb *cunnan* 'to know' and

2. Another study of grammaticization that shows how specific instances of constructions interact with more general ones is Traugott (in press). Traugott argues that certain partitive modifiers, such as *a kind/sort/bit/lot of* break off (so to speak) from the Partitive construction and realign themselves with the extant Degree Modifier construction.

the development of the Spanish auxiliary constructions that express the progressive have led us to the following more specific hypotheses:

First, prefabs are more advanced than the general construction in unithood status. As a result, the independent lexical status of the emerging gram is weakened with the effect that the gram within the prefab may be bleached of its meaning, thus contributing to the general bleaching of the meaning of the gram.

Second, in their association with semantic classes of which they are the most frequent member, prefabs promote the productivity of the general construction.

Third, a more minor tendency is that an emerging gram can be locked in a prefab, the whole of which retains an older meaning.

3. The role of prefabs in grammaticization: English *can*

Bybee (2003) traces the development of the modal auxiliary *can* from OE *cunnan* 'to know' through the end of the ME period. In OE, *cunnan* had limited use with infinitive complements; it occurred primarily with the following three classes of infinitives:

i. communication verbs, such as 'say' or 'teach', where *cunnan* meant to have the knowledge to say or teach truthfully;
ii. cognitive verbs such as 'understand', 'comprehend' or 'perceive'. As argued in Bybee (2003), these infinitives are harmonic with the 'know' meaning of *cunnan*, reinforcing it and sometimes adding more specific meaning;
iii. verbs indicating skills, again reinforcing the perhaps weakening meaning of 'know', as in 'I know the harp' by adding 'to play'.

In the Middle English (ME) texts composed by Chaucer, *can* (or *kan*) had a greatly expanded range of usage, but it continued with the same verbs and verb classes found in OE. Bybee (2003) argues that the new verbs used with *can* are related to the earlier classes of OE. In addition, in Chaucer's texts, Bybee notes some prefabs that can be identified by their relative frequency of occurrence and that the frequency of use of these tokens (such as *I can say you namoore*) may contribute to the bleaching of the meaning of *can*.

The current study investigates the latter proposal in more detail, considering the meaning of *can* in these prefabs compared to its meaning in other combinations. We find that with reporting verbs, the prefabs seem to lead to a meaning change from 'having knowledge to say' to 'being able to say' while for the cognitive verbs, where the combination of modal with main verb is harmonic, the older

usage is retained into ME and perhaps even into present day English. In this case, the older distribution is maintained, but *can* adds very little meaning.

3.1 Prefabs with 'say' and 'tell'

In the 300 tokens of *can* examined from the works of Geoffrey Chaucer, verbs of communication accounted for 102 tokens and 31 types.[3] The verbs with the highest token frequency were *tellen*, which occurred 30 times and *seye/sayn* which occurred 29 times. In general, verbs of saying and telling occur frequently in the texts because they are often used as rhetorical devices for managing the topics of the text. This is certainly true in the *Canterbury Tales*; in addition, in these tales there is often talk of who has the ability or knowledge to tell a tale and this also elevates the number of such verbs.

The following prefab with *seye* was identified on the basis of its occurrence three times in 300 tokens:

(1) I kan sey yow namoore (B. ML. 175; B. NP. 4159; G. CY. 651)

This prefab is used as a rhetorical device to end a chunk of discourse before entering another topic or scene. In this prefab, *can* indicates a notion as general as root possibility in interlocutors' interpretation of 'I can say no more because I want to get on with my narrative'. Some variations of this prefab also occur, as in (2) which omits *yow* and puts the main verb at the end:

(2) I kan no more seye (TC. 1. 1051)

Another variation uses a different negative element:

(3) I kan sey yow no ferre (A. Kn. 2060)

Another possible variation on this prefab occurs with a different verb:

(4) I kan no moore expound in this manner (B. Pri. 1725)

A different prefab shows an alternation between *sey* and *tell.* This prefab is also used as a rhetorical device to indicate the end of a portion of narrative or description. Here, however, the sense of ability is more apparent because of the adverb *bettre* which clearly points to 'ability to describe' rather than 'knowledge to say'. Note the older word order with *seye* in (5) and the word order variation with *telle* in (6). The adverb *feithfully* in (7) meant 'with faith or confidence' reinforcing the ability meaning of *kan*.

3. The tokens were the first 300 listed in Tatlock and Kennedy (1927).

(5) I kan no bettre sayn (B. ML. 42; B. ML. 874; E. Mch. 1874; I. Pars. 54)

(6) I kan telle it no bettre (B. ML. 881)

(7) I kan no bettre telle, feithfully (D. Fr. 1433)

Outside of these prefabs, a large majority of the uses of *can sey* still express the notion of 'knowledge to say', and only a few indicate ability, as indicated in Table 1.

Table 1. Other (non-prefab) uses of *kan seye* in Chaucer's English

Knowledge to say	16
Ability	3
Both	2

The two examples where both interpretations apply have the sense of 'can tell a tale', which we interpret as involving both knowledge and ability.

In comparison, the prefab uses of *can sey* do not involve knowledge to say, but are discourse devices, in the one case with a meaning of root possibility and the other a clear meaning of ability.

The situation with *can tell* is quite similar. *Telle* occurs in the prefab shown above and also in a *more than I can* + V construction exemplified by the following. One token with *telle* occurred and the others involve different communication verbs.

(8) A thousand foold wel moore than I kan telle (B. ML. 1120)

(9) And mo than I kan make of mencioun (A. Kn. 1935)

(10) And deyntees mo than I kan yow devyse (B. ML. 419)
'And dainties more than I can describe to you'

Again, this construction appears to be used as a rhetorical device for emphasizing great quantity, but the interpretation of *can* in these examples strongly suggests ability rather than knowledge.

Outside these prefabs, *can telle* is still used preferentially to express knowledge to tell, as indicated in Table 2.

Table 2. Other (non-prefab) uses of *can telle* in Chaucer's English

Knowledge to tell	15
Ability	4
Both	3

As with *seye,* the uses that allow both interpretations have as the object of *telle* a story or tale. Two of the examples that express ability are identified by the accompanying adverb and occur in a specific construction:

(11) and telle yow as pleynly as I can (A. Kn. 2481)

(12) I wol yow telle, as wel as ever I kan (A. Co. 4342)

These examples demonstrate the construction:

(13) *as* ADVERB *as* SUBJ *can*

This construction is also used with other verbs in the corpus, as shown in the following;

(14) Bot I wol passe as lightly as I kan (B. NP. 4129)

(15) As shortly as I can it trete (PF. 34)

(16) As well as that my wit can me suffyse (PF. 460)

(17) To serve you as hertly as I can (TC. 5. 941)

In these examples, the sense of *can* is clearly ability. The use of *telle* in this construction may be one of the means by which ability comes to be an interpretation of *can* with *telle*. Thus the expansion of specific constructions can be one means of spreading a new sense to a range of verb classes.

The conclusion of this section is that in the class of reporting verbs, the prefabricated or formulaic uses led the meaning change from knowledge to ability.

3.2 Cognitive verbs

Another major verb class that is used with *cunnan* in Old English contains cognitive verbs, such as *understandan, ongietan* 'understand', *tocnawan* 'to distinguish, discern' *geþenkan* 'to comprehend', and so on (Goossens 1992; Bybee 2003). As argued in Bybee (2003), these verbs are used with *can* in a way that is harmonic: the main verb echoes the meaning of *cunnan*, adding meaning that is more specific and shoring up the meaning of *cunnan* which seems to be becoming too weak to express 'knowing' on its own. These same verbs continue to occur with *can* up to the present time. Because of the harmonic nature of these expressions, *can* contributes very little to the meaning. Thus *can understand* or *can remember* are not that different in meaning from *understand* or *remember*. Indeed in most languages, no modal would be added to clauses with these verbs. Because *can* in these phrases is nearly meaningless, these expressions have likely contributed to the bleaching of *can* throughout the history of its development.

This class expands in ME as the lexicon is enhanced by borrowings from Old French. The new verbs entering the language in the 14th century come to be used with *can*. Examples found in our small corpus are: *imagine, conclude, construe, judge, remember* and *espy* (in the sense of 'discover').

3.3 Frequent items as the centers of expanding classes

We hypothesize that in the examples from the works of Chaucer the high frequency verbs are serving as the centers of the expanding classes of verbs used with *can*. This is especially clear with the two classes of verbs just discussed – the reporting verbs and the cognitive verbs. Both classes expanded greatly with the influx of lexical borrowings from Old French.

As noted above, in the Chaucer texts used, 102 of the 300 tokens were verbs of communication. There were 31 types; two of these – *say* and *tell* – accounted for 59 tokens. The evidence that these more frequent tokens serve as the central members of the category and attract other verbs with similar semantics is that some of the less frequent verbs or phrases are used in the same constructions or prefabs as *say* or *tell*. For instance, examples (9) and (10) above show *make of mencioun* and *devyce* 'describe' in a construction also used with the more frequent verbs. Of the 31 types found, 19 are verbs borrowed from Old French, suggesting that their appearance in this construction could easily have been on analogy with the other native verbs of communication that were used with *can*.

Similarly, the class of cognitive verbs found with *can* in the Chaucer texts included 18 types. The most frequent members are native English verbs – *see*, which was used in a cognitive sense nine times, *deem* and *understand* each used six times. A borrowing, *espy* 'discover', was used five times. Of the other verbs and expressions in this class, ten were borrowed from Old French. Since we have argued that the origins of *can* with cognitive verbs is an harmonic construction, it follows that the new verbs and expressions were used with *can* on analogy with the established, and more frequent, verbs in this construction.

4. Spanish progressive and other imperfective gerund periphrases

The development of a set of progressive constructions from Old Spanish to Modern Spanish provides us with the opportunity to study the structural as well as semantic properties of grammaticizing constructions and their conventionalized instantiations.

4.1 The grammaticization of the progressive in Spanish

In Old Spanish (12th – 15th centuries) texts we find occurrences of a general gerund construction, in which finite forms of spatial (locative, postural, or motion) verbs combine with another verb in gerund (*-ndo*) form to mean 'be/go VERB-ing', as shown in (18):

(18) Gerund construction: [Verb $_{\text{locative-postural-motion}}$ + gerund (*-ndo*)] $_{=\text{'be/go VERB-ing'}}$

The verbs occurring in the finite verb slot are:

(19)

Location-Postural	Movement
estar 'be (located)'	*andar* 'go around'
quedar 'remain, stand still'	*ir* 'go'
yacer 'lie'	*salir* 'go out'
	venir 'come'

The finite form is an independent lexical item with full spatial meaning, as illustrated in the 13th c. examples in (20–22). Lexical status is indicated by a co-occurring locative, which may (as in [20], 'in that pond') or may not (as in [21], 'along the road') intervene between the finite form and the gerund. Lexical status is also evident in the combination of a motion verb with another motion verb in a harmonic use, where the gerund describes the manner of motion (as in [21], 'go (by) walking'). Finally, in (22), the repetition of *andar* 'go around' and *buscar* 'look for' separately shows that *andando buscando* is a combination of two independent lexical items.

(20) Et alli ESTAUA el puerco en aquella llaguna BOLCANDO se (XIII, GE.II)
'And there WAS the pig in that pond TURNING itself'

(21) YUASSE ANDANDO por la carrera que ua al pozo (XIII, GE.I)
'He WENT WALKING along the road that goes to the well'

(22) Et ANDANDO BUSCANDO los. encontrosse con un omne quel pregunto como andaua o que buscaua. (XIII, GE.I)
'And GOING AROUND LOOKING FOR them he met a man who asked him how he was going or what he was looking for'

Particular instances of this general gerund construction grammaticize, yielding a set of aspectual constructions (cf. Bybee 2006). Thus, *estar* 'be located', *ir* 'go', and *andar* 'go around' + V-*ndo* evolve from lexical spatial expressions into grammatical aspectual morphemes in these constructions. In present-day varieties of Spanish, these gerund periphrases cover a range of meanings in the domain of imperfective aspect (e.g., Camus Bergareche 2004). In particular, the construction *estar* + V-*ndo* as shown in (23) is on its way to becoming an obligatory expression of progressive aspect in the Present tense (Torres Cacoullos 2000, Chapter 5; García Fernández et al. 2006: 140).

(23) [*Estar* + V-*ndo*] $_{=\text{progressive}}$

Throughout the evolution of these gerund periphrases, there is retention of spatial meaning from the source construction (Bybee & Pagliuca 1987; Hopper 1991) and spatial and aspectual meanings coexist synchronically, often in the same token. For example in (24), from a corpus of New Mexican Spanish, *está cuidando*

televisión means both 'he is there, in front of the TV' (locative – lexical) and 'he is in the midst of an activity at reference time, i.e., watching TV' (progressive – grammatical) (Torres Cacoullos 2000: 9).

(24) *- ¿Aquí está?*
- Sí ESTÁ CUIDANDO *televisión.*
- Oh.
- Ahi en en en el cuarto allá del otro lado. Está dormido en la silla. (NMbil/Vig)
'- Is he here?
– Yes he IS WATCHING television.
– Oh.
– There in in in the room over there on the other side. He's asleep in the chair'

At the same time, aspectual meaning is present from the earliest texts. In the next set of 13th c. examples, locative or physical motion meaning is less discernable than in (20–22), rather the meaning is more aspectual, with *estar* + V-*ndo* indicating a situation in progress (25), *ir* + V-*ndo* a gradually developing process (26), and *andar* + V-*ndo* figurative motion together with continuous meaning (27).

(25) cato por una finiestra & uiol estar con ella […] como ESTA marido FABLANDO con su muger (XIII, GE.I)
'he looked through a window and saw him (be) there with her […] as IS a husband SPEAKING with his wife'

(26) porque non poblara el y […] & YUAN ya las yentes SEYENDO muchas. (XIII, GE.I)
'so that he wouldn't settle there […] and the people already WERE GROWING' (literally: went the people being many)

(27) el que […] quiere andar los caminos peligrosos ANDA BUSCANDO su muerte (XIII, Calila)
'he who [..] wants to walk dangerous roads IS LOOKING for his death'

4.2. Grammaticization indices for Spanish progressives: Locatives and unithood

Grammaticization of the finite locative-motion verb in gerund periphrases proceeds via semantic reduction, which in this case involves the loss of spatial meaning (Torres Cacoullos 2000: 71–113). Yet we cannot establish that grammaticization is occurring by comparing isolated examples from earlier and later periods, since throughout the evolution of gerund periphrases there is retention of spatial meaning from the source construction, even in present-day examples (such as [24]). Nor would quantitative comparisons across periods of the proportion of tokens with aspectual as opposed to spatial meaning be a replicable measure, since tokens may be compatible with both meanings (again as in [24]) and analysts' interpretations may

well differ. Instead, we can show the advance of grammaticization by uncovering changes in distribution patterns.

Tokens of the gerund construction were exhaustively extracted from 13 texts, representing four periods: late 13th c. (three texts, approximately 900,000 words), late 15th c. (five texts, approx. 500,000 words), early 17th (one text, approx. 400,000 words), late 19th (four texts, approx. 350,000 words) (see Corpus, before References, and Table 6 for token counts).[4] We first present distribution patterns for *estar* 'be located' and then some results for *ir* 'go' and *andar* 'go around' + V-*ndo*.

We hypothesize that bleaching of spatial meaning will be shown in a decrease of co-occurring locatives, in the aggregate.[5] Table 3 shows the percentage of *estar* + V-*ndo* tokens with a co-occurring locative, in the four chronological sets. The rate of co-occurring locatives diminishes, from an average of 38% (91/238) in the 13th and 15th c. (Old Spanish) data combined, to 24% (51/217) in the 17th c. and 16% (35/217) in the 19th c. data. We take this result as a measure of loss of spatial meaning and thus advancing grammaticization.

Table 3. Co-occurring locatives in Progressive *estar* + *V-ndo*

XIII	XV	XVII	XIX
36% (37/104)	40% (54/134)	24% (51/217)	16% (35/217)

XIII-XV 38% (91/238) vs. XVII-XIX 17% (86/434) Chi-Square 15.10291903; p = 0.0001.

A second measure of the grammaticization of Progressive *estar* + V-*ndo* and the motion-verb (*ir, andar*) + V-*ndo* periphrases is the degree of unithood. Bybee (2003: 603) proposes that frequent collocations become automated as single processing units, gaining autonomy in two ways. Analyzability is lost when the erstwhile individual constituents of the frequent collocation weaken their association with other instances of the same constituents and with other instances of the same construction. We examine three indices of unithood: adjacency, association, and fusion (Torres Cacoullos 2000, Chapter 2).

4. The texts are chronicles (13th and 15th c.) and novels; the 15th c. corpus includes two plays (the *Celestina* and the early 16th c. *Lozana*).

5. Locative co-occurrence need not always indicate a lesser degree of grammaticization; the locative may promote the aspectual meaning of the auxiliary when it refers to the main verb in harmonic uses (for example, *ir* plus another motion verb) (cf. Hopper & Traugott 1993: 83) or when it is incompatible with the auxiliary's original spatial meaning (for example an allative locative with *estar*).

1. Adjacency: the locative-motion finite verb and the gerund may be adjacent or they may be separated by intervening material. In the 13th c. data, nearly two thirds of *estar* + V-*ndo* tokens have an intervening locative or temporal adverbial, subject or object, or a combination of elements, as in (28) and (29).[6]

(28) está Melibea muy affligida HABLANDO con Lucrecia sobre la tardança de Calisto (XV, Celestina, XIV, 282)
'[Stage instructions] is Melibea, deeply distressed, TALKING to Lucrecia about the tardiness of Calisto' (cf. Singleton, 197)

(29) Pero hombre, ¿estamos locos? … ¿qué está usted HABLANDO? (XIX, Perfecta, 284)
'But man, are we crazy? … What ARE you TALKING about?

2. Association: multiple gerunds may co-occur, as in (30), or the finite verb may be more tightly associated with a single gerund as in (31), where *ir* is repeated for each gerund.

(30) le YVAN MENGUANDO los bastimentos e CRECIENDO las necesidades (XV, CRC LIV, 178)
'supplies WERE [lit: went] SHRINKING and needs GROWING'

(31) la vida vulgar VA PENETRANDO y se VA INFILTRANDO en mi naturaleza. (XIX, Pepita, 55)
'ordinary life IS (gradually) [lit: goes] PENETRATING and IS [lit: goes] INFILTRATING my nature'

3. Fusion: object pronouns may appear as enclitics on the gerund or proclitics on the finite verb. This latter configuration, called "clitic-climbing" (e.g., Myhill 1988), is a manifestation of greater fusion between the emergent auxiliary and the gerund: in (32), proclitic *los* indicates that *fueron conservando* is a unit, just like single-word *conservan*.[7]

(32) otros, que tuvieron principios grandes, y LOS FUERON + CONSERVANDO y los conservan y mantienen en el ser que comenzaron; (XVII, Quijote II, VI)
'others had noble beginnings, and PRESERVED [lit: went preserving] **them**, and still preserve and maintain them just as they were' (Grossman, 494) [lit: went preserving, i.e., continued (went on) preserving them]

---

6. In counting *estar* + V-*ndo* tokens, we included cases of intervening adjectives (N=35) (but not *estarse quedo* + V-*ndo* in the Quijote, N=7); though it could be argued that *estar* + Adjective + V-*ndo* is a different construction, it does not exclude progressive meaning and thus is associated with the more general *estar* + V-*ndo* construction.

7. Excluded from the count were cases of structurally ambiguous reflexive marking, which may have contributed to the increase of clitic climbing over time (Torres Cacoullos 2000: 50–51).

Table 4 shows a diachronic increase in adjacency, association, and fusion for *estar* + V-*ndo*. The proportion of occurrences without intervening material increases significantly between all the data sets (from 36% in the 13th, to 50% in the 15th, 67% in the 17th, and 78% in the 19th century). The proportion of occurrences with a single as opposed to multiple gerunds increases from 80% in the 13th to 92% in the 19th c. data. And the rate of "clitic climbing" shows an increase between the combined 13th and 15th c. data, at 57%, and the combined 17th and 19th c. data, at 76% (we attribute the later drop in rate to the development of stylistic meaning in clitic climbing in the 19th c. (Torres Cacoullos 1999)).

Table 4. Grammaticization (unithood) measures for *estar* + V-*ndo*: Adjacency (lack of intervening material), Association (absence of multiple gerunds), Fusion ("clitic climbing")

	XIII	XV	XVII	XIX
Adjacency	36% (37/104)	50% (67/134)	67% (145/217)	78% (169/217)
Association	80% (83/104)	86% (115/134)	88% (192/217)	92% (199/217)
Fusion	63% (15/24)	50% (11/22)	82% (61/74)	70% (54/77)

Adjacency: XIII vs. XV Chi-Square 4.950998521; p = 0.0261; XV vs. XVII Chi-Square 9.799123895; p = 0.0017; XVII vs. XIX Chi-Square 6.634394904; p = 0.0100. Association: XIII vs. XIX: Chi-Square 9.323668501; p = 0.0023. Fusion: Combined XIII-XV vs. XVII-XIX: 57% (26/46) vs. 76% (115/151) Chi-Square 6.682716664; p = 0.0097

Based on these three unithood indices, we constructed a cumulative "grammaticization index", weighted to take account of adjacency more than association and fusion, as follows:

Adjacency: two points for no intervening material, one for an intervening subject, object, temporal or manner expression, zero for an intervening adjective, locative or more than one of the above.

Association: one point for a single as opposed to multiple gerunds.

Fusion: one point for a proclitic as opposed to enclitic. Since clitic climbing does not apply to all tokens, the index is calculated as a fraction.

Table 5 shows a diachronic increase in the value of this index for *estar* + V-*ndo*.

Table 5. Cumulative grammaticization (unithood) index for estar + V-ndo

XIII	XV	XVII	XIX
.60 (62.2/104)	.74 (99.33/134)	.79 (172.5/217)	.83 (180.75/217)

*Between parentheses is the point total for all tokens divided by the number of tokens

A final measure of the advancing grammaticization of *estar* + V-*ndo* is relative frequency. Table 6 shows the changing relative frequency of the locative-motion verbs in gerund periphrases. From having half the relative frequency of

ir in the 13th c. data, *estar* goes on to become the most frequent in the cohort of emerging auxiliaries; *ir* + V-*ndo* remains viable, but its frequency relative to *estar* decreases; and *andar* + V-*ndo* ceases to be productive (in these Peninsular Spanish data).

Table 6. Relative frequency of gerund (V-*ndo*) periphrases

	XIII (N=477)	XV (N=301)	XVII (N=505)	XIX (N=557)	
estar 'be'	26%*	45%	43%	39%	Rises
ir 'go'	50%	27%	37%	35%	Decreases
andar 'go around'	21%	16%	12%	3%	Declines
venir 'come'	3%	9%	4%	3%	Always minor
seguir 'follow'	1	1	0	15%	Appears late
quedar 'remain'	0	3%	4%	2%	Always minor
continuar 'continue'	0	0	0	3%	Appears late

The relative frequency of *estar* is greater in the XV c. than in the XIII c. (Chi-square 29.99123288; p = 0.0000); differences in the relative frequency of *estar* between the XV c., XVII c., XIX c. are not significant.

*13th c. *estar* count includes 18 tokens of *seer* + V-*ndo*.

In summary, *estar* + V-*ndo* shows bleaching of locative meaning (Table 3), an increasing unithood index (Tables 4, 5), and increasing relative frequency (Table 6). In the next section we examine the role prefabs have played in the grammaticization of *estar* + V-*ndo*.

4.3 Prefabs and *estar* + V-*ndo* grammaticization

In identifying prefabs, we consider relative frequency rather than token frequency, both with respect to the "auxiliary" and the gerund (cf. Torres Cacoullos 2000: 57–59, 2006; Hay 2001). We operationally define prefabs as "auxiliary"-plus-gerund combinations making up 2% or more of the corresponding "auxiliary" data and 50% or more of the corresponding gerund data. For example, *estar hablando* 'be talking' makes up 5% (32/672) of *estar* data and 71% (32/45) of *hablando* data. Combining the data of all time periods, we identified the prefabs appearing in Table 7 (listed alphabetically, by "auxiliary").[8]

8. *Estar hablando* total includes four 13th c. tokens of *seer hablando*. High frequency *diciendo* 'saying, telling' (N=50), which makes up 3% (22/672) of the *estar* and 2% (15/700) of the *ir* data, is not overwhelmingly associated with either auxiliary (44% (22/50) *estar*, 30% (15/50) *ir*).

Table 7. Prefabs (as percentage of aux and of gerund; all time periods combined)

		% "auxiliary"	% gerund
ESTAR	*aguardando* 'waiting'	2% (14/672)	93% (14/15)
	diciendo 'saying'	3% (22/672)	44% (22/50)
	durmiendo 'sleeping'	2% (14/672)	93% (14/15)
	escuchando 'listening'	3% (23/672)	96% (23/24)
	esperando 'waiting'	7% (48/672)	89% (48/54)
	hablando 'talking'	5% (32/672)	71% (32/45)
	mirando 'looking'	7% (49/672)	84% (49/58)
	oyendo '*hearing*	2% (15/672)	94% (15/16)
	pensando 'thinking'	2% (13/672)	62% (13/21)
IR	*creciendo* 'growing'	3% (24/700)	86% (24/28)
	diciendo 'saying'	2% (15/700)	30% (15/50)
	entrando 'entering'	3% (22/700)	100% (22/22)
	haciéndose 'becoming'	3% (24/700)	100% (24/24)
	huyendo 'fleeing'	3% (22/700)	67% (22/33)
	yendo 'going'	2% (14/700)	88% (14/16)
	llegando 'approaching, arriving'	3% (20/700)	95% (20/21)
	viniendo 'coming'	2% (12/700)	92% (12/13)
ANDAR	*buscando* 'looking for'	25% (57/229)	84% (57/68)
VENIR	*huyendo* 'fleeing'	13% (10/77)	
SEGUIR	*andando* 'walking'	5% (4/85)	
	creciendo 'growing'	4% (3/85)	
	siendo 'being'	5% (4/85)	
QUEDAR	*esperando* 'waiting'	12% (5/43)	

We will make a case that (1) prefabs are in the advance of the general construction in unithood status in early stages and thus demote the independent lexical status of the emerging auxiliary, and (2) in their association with semantic classes of which they are the most frequent member, prefabs promote the productivity of the general construction.

4.3.1 *Prefabs lead as units*

The first column of Table 8 shows *estar* + V-*ndo* prefabs by time period. Two prefabs in particular, *estar hablando* 'be talking' and *estar esperando* 'be waiting', are evident throughout the time periods examined and continue in present-day data. *Estar hablando* is the single most frequent *estar* + V-*ndo* collocation (165/2270) in conversational Peninsular Spanish data (COREC, Marcos Marín 1992) and *estar esperando* (38/2270) is still among the top ten collocations.[9]

9. *Haciendo* 'doing' is more frequent than *hablando* in the COREC data (N = 216), but it combines with (often non-referential) objects to form different predicates, thus we don't view it as a single collocation like *estar hablando*.

Table 8. Estar + V-*ndo* prefabs, by time period: Comparison of grammaticization indices

		Unithood index		Locative	
		prefab	**general estar + V-ndo (Table 5)**	**prefab**	**general (Table 3)**
XIII	*hablando*	.67 (8/12)	.60 (62.2/104)	17%	36%
	esperando	.82 (4.9/6)		33%	
XV	*hablando*	.89 (8/9)	.74 (99.33/134)	11%	40%
	esperando	.74 (13.3/18)		56%	
XVII	*hablando*	.72 (4.33/6)	.79 (172.5/217)	33%	24%
	diciendo	.90 (9/10)		0	
	mirando	.79 (29.83/38)		5%	
	escuchando	.79 (14.25/18)		6%	
	esperando	.78 (16.33/21)		38%	
	Avg	.79 (73.74/93)			
XIX	*hablando*	.93 (4.67/5)	.83 (180.75/217)	20%	16%
	diciendo	1.00 (4/4)		25%	
	mirando	.73 (4.42/6)		0	
	oyendo	.76 (6.83/9)		22%	
	pensando	.96 (8.67/9)		0	
	esperando	.75 (2.25/3)		67%	
	Avg	.86 (30.84/36)			

As with English can, gerund construction prefabs may begin as harmonious expressions, where the original lexical meaning of the emerging auxiliary is compatible with the main verb. That is, as Torres Cacoullos (2000: 175) has argued, frequent collocations such as *estar hablando* 'be talking', *ir creciendo* 'be (go) growing' and *andar buscando* 'be (go around) looking for' (Table 7) "follow from the original uses of the source constructions". Such harmonious prefabs may appear conservative in manifesting retention of meaning from the source construction, for example, a locative meaning component in *estar esperando* 'be waiting', as in (33): over one-third of present-day oral Peninsular Spanish (COREC) tokens (34%, 13/38) have a co-occurring locative, whereas the rate of co-occurring locatives with *estar hablando* is 5% (9/165) in the same corpus. Nevertheless, retention of original meaning in the unit, originally a harmonic combination, does not detract from grammaticization. On the contrary, since the locative meaning is contributed by *esperar* 'wait', the meaning contribution of *estar* is minimized.

(33) y él nos ESTABA ESPERANDO en San Sebastián (COREC, CCON035B)
'and he WAS WAITING for us in San Sebastián'

Other prefabs may conventionalize as fixed discourse formulas. For example, *estoy hablando de* 'I'm talking about' or *estamos hablando de* 'we're talking about',

as in (34), (see also [29]) may play more of an interactional role akin to discourse markers or connectives rather than actually referring to a situation in progress. Some scholars call such developments "pragmaticalization" (e.g., Erman & Kotsinas 1993; cf. Aijmer 1997: 3). As with prefabs manifesting meaning retention, prefabs with formulaic discourse uses detract from the independence and meaning contribution of the erstwhile lexical item (locative or motion verb).

(34) Estamos hablando de la madre no del matrimonio. (COREC, PEDU010A)
'We're talking about the mother not the couple'

Both these prefabs show an early lead in their unithood index. The columns in Table 8 compares the unithood indices and rate of co-occurring locatives for the prefabs and the general construction (all tokens of *estar* + V-*ndo*). *Estar hablando* leads the grammaticization of *estar* + V-*ndo* in the earliest (Old Spanish) stage, with a unithood index of .67 compared to .60 for the general construction, in the 13th c., and .89 compared to .74, in the 15th c. data. The rate of cooccurring locatives is also lower with *estar hablando*, at 17% and 11%, compared to 36% and 40%, in the 13th and 15th c. data, respectively. *Estar esperando* also shows a higher than average unithood index, in the 13th c. data, though not a lower locative rate. Over time, as the productivity of the general construction increases, *estar hablando* makes up a smaller portion of the data, from 12% (12/104) of all *estar* + V-*ndo* tokens in the 13th c. to 2% (5/217) in the 19th c., and appears to follow general patterns.

Thus, in early stages, prefabs score higher than the general *estar* + V-*ndo* construction on the unithood measures shown above (Section 4.2). This empirical result provides evidence that frequent collocations become automated as single processing units (Bybee 2003). As we argued earlier, prefabs contribute to grammaticization because they are accessed holistically, which means that the erstwhile independent lexical item contributes less meaning, which promotes the semantic bleaching of the emerging auxiliary in this construction. Thus, it is the unithood of prefabs, meaning retention or formulaic discourse uses notwithstanding, that is conducive to grammaticization.

Now, given the relative autonomy of high frequency collocations (Bybee 2003), how do these prefabs contribute to the productivity of a general grammatical construction? Our argument is that prefabs maintain associations with the more general construction. In the next section we will show that prefabs contribute to productivity via the semantic classes centered around them.

4.3.2 *Prefabs contribute to productivity via associated semantic classes*

Estar + V-*ndo* prefabs are *estar hablando* and *estar esperando* in the Old Spanish data, as we have seen; these plus *estar diciendo* 'be saying, telling', *estar mirando* 'be watching', *estar escuchando* 'be listening' in the 17th c. data; and all of the above plus *estar pensando* 'be thinking' and substituting *oyendo* 'hearing' for *escuchando* 'listening' in

the 19th c. data. From Table 8 we can deduce the proportion of the data made up by the prefabs by adding their tokens (the second number between parentheses in the first column) and taking this sum over the total number of tokens per time period (the second number between parentheses in the second column). This proportion seems to remain steady over time at 17% (18/104) in the 13th c. data, 20% (27/134) in the 15th, 43% (93/217) in the 17th (38 tokens of *estar mirando* 'staring or gazing' in the *Quijote* contribute to this inflated figure), and 17% (36/217) in the 19th c. data. Nevertheless, considering we have listed six instead of two prefabs in the 19th c., it is fair to conclude that while there is continuity of particular prefabs over time, these make up a declining proportion of the general construction data. This is as expected, since grammaticization involves generalization to more and more types.

The prefabs participate in classes with other semantically related verbs, ranging from the large class of verbs of speech (e.g., *alabar* 'praise', *demandar* 'request', *explicar* 'explain', *gritar* 'shout', *murmurar* 'murmur', *razonar* 'argue', *rogar* 'beg, pray') to the small class of verbs of 'waiting'. Intuitively apparent semantic classes for the prefabs identified in Table 8 are shown in (35). Besides noting the verbs of speech and 'waiting' verbs, we coded all tokens for affiliation with verbs of perception, bodily activity (e.g., *bañarse* 'bathe', *doler* 'ache', *llorar* 'weep', *respirar* 'breathe', *sangrar* 'bleed', *temblar* 'tremble'), and cognition-emotion (e.g., *figurar* 'imagine, think', *morirse de miedo* 'be scared to death', *penar* 'suffer', *rumiar* 'ruminate', *sentir* 'feel', *temer* 'fear'). Table 9 shows the distribution of *estar* + V-*ndo* tokens in semantic classes, by time period. The distribution and concentration of tokens in the semantic classes we defined appears steady (we will return shortly to the decline of the 'waiting' class).

(35) *Estar*: Prefabs (Table 8) and semantic classes

hablando, diciendo	SPEECH class size:	big
pensando	COGNITION (also emotion)	big
durmiendo	BODY ACTIVITY	medium sized
escuchando, mirando, oyendo	PERCEPTION	small
esperando, aguardando	WAITING	very small

Table 9. Semantic classes: *Estar* + V-*ndo*

	XIII (N=104)	XV (N=134)	XVII (N=217)	XIX (N=217)	
Speech	16% (17)	18% (24)	13% (28)	15% (32)	steady
Cognition	11% (11)	7% (9)	4% (9)	12% (27)	steady
Body activity	13% (13)	4% (5)	8% (17)	11% (23)	steady
Perception	12% (12)	7% (10)	28% (60)*	10% (22)	steady
Waiting	11% (11)	15% (20)	13% (29)	3% (6)	decline
Other	38% (40)	49% (66)	34% (74)	49% (107)	steady

*In Quijote, *estar escuchando* N=18, *estar mirando* N=38

** Difference proportion "Other" XIII-XV combined 45% (106/132) vs. XVII-XIX combined 42% (181/253) is not significant.

Though contrary perhaps to our expectations we do not see an increase in the "Other" category, that is, an expansion outside the original semantic classes over time, generalization of *estar* + V-*ndo* is shown in a count of type/token ratios, where "types" are the different verbs appearing in the open slot in the construction. Table 10 shows type/token ratios, calculated for each data set based on a random sample of 100 tokens (since an increased sample size is likely to show a lower type/token ratio, as lexical types are repeated). The ratio increases from 48–49 in the 13th and 15th c. data, to 55 in the 17th and 69 in the 19th c. The increase in type/token ratio over time indicates the increased productivity of the construction but the lack of increase in the "other" category in Table 9 indicates that much of the generalization is taking place within the established verb classes.

Table 10. Type/token ratio: *Estar* + V-*ndo* (randomized sample 100)

XIII	XV	XVII	XIX
49/100	48/100	55/100	69/100

Even though *estar hablando* and *estar esperando* are both high frequency prefabs, an important difference is precisely that the former is part of the large class of verbs of speech appearing in the *estar + V-ndo* configuration, while the class of verbs of 'waiting' is tiny, including only *aguardar* and *atender* besides *esperar* ([35]). Since *estar hablando* is associated with a high type frequency semantic class, the contribution of this prefab to the development of a general *estar* + V-*ndo* construction should be greater than that of *estar esperando*. Besides the striking decline in the relative frequency of 'waiting' verbs (shown in Table 9), two pieces of evidence show the weaker contribution of *estar esperando* to the grammaticization of the general construction. First, recall that *estar esperando* has had a higher than average rate of co-occurring locatives from the 15th c. data onwards (Table 8). Second, though *esperando* is still among the top ten or so gerunds combining with *estar* (in the present-day COREC data), its exclusive association has eroded. While in the 13th and 15th c. data, 100% (24/24) of *esperando* tokens co-occurred with *estar* as opposed to another "auxiliary", beginning with the 17th c. data, *quedar* 'remain' combines with this gerund, so that *quedar esperando* is somewhat of a prefab (by our operational definition) in its own right, making up 12% (5/43) of all *quedar* + V-*ndo* tokens (Table 7). In contrast, no other "auxiliary" competes with *estar*'s association with *hablando*. So as predicted, the contribution of *estar esperando* and its low type frequency class to the grammaticization of a general *estar* + V-*ndo* construction is less consistent than that of *estar hablando*.

The conclusion of this section is that prefabs may participate in classes with other semantically related verbs and that these classes may be higher or lower type frequency categories. We reason that participation in high type frequency categories, as in the case of *estar hablando*, contributes to a more general schema and thus greater productivity (Bybee & Eddington 2006; cf. Torres Cacoullos 2000: 13, 130). In contrast, if the prefab cannot be associated with a many-membered class, as is the case with *estar esperando*, it will not contribute as consistently to the productivity of the grammaticizing construction.

4.3.3 *Prefabs and productivity: Evidence from motion-verb gerund periphrases*
Further support for the hypothesis that prefab exemplars of a grammaticizing construction must be associated with semantically related instances in order to contribute to the productivity of the construction is provided by *ir* + V-*ndo* and *andar* +V-*ndo* distributions.

Ir + V-*ndo* has developed a meaning of 'gradually developing' or prospective imperfective aspect (cf., e.g., Dietrich 1983; Olbertz 1998; Squartini 1998). The data suggest that this more general construction emerges from more particular *ir* + V-*ndo* constructions, including a harmonic motion construction and a change-of-state construction (Torres Cacoullos 2000: 151). One set of *ir* + V-*ndo* prefabs in the early data is harmonic motion expressions with *yendo* 'going', *llegando* 'arriving, nearing', *viniendo* 'coming'; another prefab set is process verb expressions *ir creciendo* '(gradually) grow' and *ir haciéndose* '(gradually) become' (Table 7, above). The two corresponding semantic classes, motion verbs and process (change-of-state) verbs, which include many other members, have been the mainstay of the construction, making up between one-third and one-half of all the *ir* + V-*ndo* data in all time periods, as shown in Table 11. While the proportion of motion verbs has declined, as expected if the construction has grammaticized from a harmonic motion verb expression, process verbs appear to remain stable. A measure of the association of *ir* + V-*ndo* with processes is cooccurrence with reflexive (*se*)-marked lexical types, a number of which refer to changes of state, for example, *mudarse* 'change', *tornarse* 'become', and which pair up with *ir* as opposed to *estar* (though *estar* + V-*ndo* has generalized even to this context).[10]

10. The ratio of *ir* + $V_{\text{REFLEXIVE}}$ *–ndo* to *estar* $V_{\text{REFLEXIVE}}$ *–ndo* tokens shows a decline: 13th c. 48: 7 > 15th c. 10: 5 > 17th c. 23: 13 > 19th c. 38: 23.

Table 11. Semantic classes: Ir + V-*ndo*

	XIII (N=238)	XV (N=80)	XVII (N=188)	XIX (N=194)	
Motion	37% (88)	23% (18)	29% (55)	23% (44)	decline
Process	18% (43)	11% (9)	6% (12)	20% (38)	steady
All other verbs	45% (107)	66% (53)	64% (121)	58% (112)	

Though it starts out with more than double the relative frequency in 13th c. data, *ir* + V-*ndo* is not as productive as *estar* + V-*ndo.* Over time it is overtaken by *estar* + V-*ndo* in relative frequency (Table 6) and the pace of grammaticization has been slower for *ir* + V-*ndo*, as indicated in Table 12: we find no significant decrease in co-occurring locatives and two of the three unithood indices, association (single vs. multiple gerunds) and fusion (clitic climbing) fail to show an increase (adjacency, that is, lack of intervening material, does increase, from 58% (137/238) in the 13th c. data to 89% (172/194) in the 19th c. data (Chi-Square 50.749219454; p = 0.0000)). Furthermore, some of *ir* + V-*ndo*'s uses have been taken over by newcomer (in the 19th c.) *seguir* 'follow, continue' + V-*ndo*, at least in some varieties (Tables 6 and 7).

Table 12. Grammaticization indices *ir + V-ndo*: Co-occurring locatives and unithood measures

	XIII N = 238	XV N = 80	XVII N = 188	XIX N = 194
Locatives	25%	24%	19%	21%
Adjacency	58%	61%	86%	89%
Association	85%	84%	90%	90%
Fusion	95%	88%	95%	66%

How do we explain the restricted productivity of *ir + V-ndo* compared to *estar* + V-*ndo* despite an early lead in relative frequency (Table 6)? Contributing to grammaticization is the persistence of early prefabs (such as *ir creciendo* 'be [go] growing') and their association with high type frequency semantic classes participating in the construction (such as the process verb class). At the same time, however, from the beginning the construction has been heavily concentrated in a small number – only two – classes due to its more specific meaning, in contrast to the more general *estar + V-ndo*, which has been more evenly distributed across different semantic classes (Table 9).

The single-most remarkably robust prefab is *andar buscando* 'be [go] looking', which makes up an average of 20% of the tokens of the *andar* construction.

Andar buscando is so frequent that it has been said to be "really a set phrase" (Spaulding 1926: 259) or a case of "lexical specialization" (Squartini 1998: 261). Just like *estar hablando* and *ir creciendo, andar buscando* continues as a well-established routine in present-day varieties (with 9% (8/89) of all *andar* + V-*ndo* tokens in a corpus of popular Mexican Spanish (Torres Cacoullos 2000: 168)). But unlike *estar hablando* and *ir creciendo*, this prefab is not associated with a large class of semantically related items (though other lexical types in the Old Spanish data take on a 'looking for' meaning, for example *andar demandando* 'enquiring' and *andar catando manera* 'looking for a way' or *andar guisando cómo* 'arranging how' (Torres Cacoullos 2000: 164–165)). As we would predict, *andar* + V-*ndo* shows a sharp decline, dropping from 21% in the 13th c. to a relative frequency of 3% in the 19th c. data. Social factors are clearly important, since *andar* + V-*ndo* is much more frequent in other varieties, especially Mexican Spanish, where it has developed social associations (Torres Cacoullos 2001). Nevertheless, the restriction of *andar* + V-*ndo* compared to *ir* + V-*ndo* and especially *estar* + V-*ndo* is consonant with the notable strength – and isolation – of its prefab.

In summary, *ir* + V-*ndo* remains a viable aspectual expression, though largely concentrated in two semantic classes, while *andar* + V-*ndo* is geographically and socially restricted. Both the viability of *ir + V-ndo* and its slower grammaticization as well as the restriction of *andar + V-ndo* would be predicted by the view of prefabs and their associated semantic classes that we are advancing: early prefabs persist but contribute to productivity (generalization) of a grammaticizing construction only if they are associated with relatively large semantic classes of lexical types participating in the construction.

5. Conclusions

Our study, then, contributes to the understanding of the relation between the specific and the general in the development of constructions over time. We hope to have shown that prefabs are important to the understanding of the fabric of grammaticization. At any given point in time, prefabs will be responsible for increasing the frequency of grammaticizing constructions as well as for serving as the loci for extensions of the construction. Their lack of compositionality, their frequency and conventionalization play an important role in providing meaning for the construction as a whole while at the same time affecting the meaning of the constituent parts, usually by loss of earlier, lexical meanings. These interactions demonstrate that prefabs and their related constructions remain associated and interact in language change.

Specifically, we have demonstrated that prefabricated instances of constructions lead in the semantic reduction of the meaning of the construction as well as in manifesting structural indices of unithood. As predicted from their relative frequency of use, prefabs grammaticize earlier or at a faster rate than the general construction.

We have also presented evidence that prefabricated instances of constructions serve as the centers of subclasses of the grammaticizing construction, attracting more lexical types into the construction and thereby contributing to the productivity of the construction. This process is apparent in the Middle English verbs of communicating and cognition with *can,* as well as in the verbs of communicating with the *estar-* and in the process verbs with the *ir-* progressive construction in Spanish.

To a lesser extent, we find prefabs retaining the older meaning or distribution of the construction. As we mentioned, *estar esperando* 'to be waiting' still often co-occurs with locative expressions, suggesting that this exemplar implies 'waiting somewhere'. We argue that it is not so much that *estar* has retained its locative meaning here as that the whole prefab has a locative implication that derives as much from *esperar* as from *estar.* Similarly in English, the expression *I kan nought sayn* from Middle English and its modern descendent, *I can't say*, in some uses gives a knowledge interpretation: 'I cannot say because I don't know'. Again, the parts of the construction are harmonic in that saying itself implies knowledge to say. Thus it is the whole prefab that retains the earlier meaning, not just the auxiliary.

In other cases, an older distribution is maintained by a prefab, while the older meaning has eroded. Thus *andar buscando* 'to be looking for' is purely aspectual, but the use of *andar* with *buscar* reflects an older compatibility of the two lexical items. In the English examples, we have the continued use of *can* and *can't* with main verbs such as *understand, remember, imagine, guess, believe*, where the modal contributes very little if any meaning. The use of *can* with these cognitive verbs is retained from the very earliest period when *cunnan* meaning 'know' was harmonic with these more specific verbs.

Our study has both diachronic and synchronic implications. To come back to the dimensions along the continuum between prefabs and more general constructions that we presented in the introduction, the data we have examined shows how essential it is that we consider prefabs to be highly integrated with the more general constructions.

> Productivity: Even within a general construction, such as *can* + VERB or *estar* + VERB *–ndo,* there can be expressions with varying degrees of productivity: *can* + cognitive verb occurs with many different types, as does *estar* + speaking verb, while *estar* with *esperando* is quite isolated.

Transparency of meaning: *can't say* retains the knowledge interpretation while *can't understand* has no real semantic role derivable from *know* for *can,* yet these are clearly instances of the same construction. Also, *estar esperando* retains some locative nuance, while other instances of *estar* + gerund have lost all such meaning.

Analyzability: As the Spanish data show, the degree of analyzability can also vary, with more frequent collocations showing less analyzability as demonstrated by less frequent occurrence of modifiers and multiple gerunds and the more frequent occurrence of proclitics before the whole expression.

Thus it appears that grammaticization of a construction is not a uniform process with all instances or subclasses of the construction marching through the changes in lockstep. Rather, certain instances of the construction lead the charge, attracting other similar expressions, while low frequency uses may drag along at the rear. Some high-frequency instances may become fossilized early on, maintaining older meanings, while others rush ahead to become bleached and generalized. Our more general point, then, is that prefabs are not marginal or peripheral to grammar at all, but rather highly integrated with the more general structures of the language. Thus language use with its varying lexical specificity and uneven contours of token and type frequency is highly involved in the creation and maintenance of grammatical constructions.

Corpus [word counts-tokens]

Calila (1250) = Anonymous. 1987. *Calila e Dimna*, ed J.M. Cacho Blecua and M.J. Lacarra. Madrid: Castalia. [86,000–30]

GEI (1260–1280) = Alfonso X. 1930. *General estoria. Primera parte*, ed. Antonio G. Solalinde. Madrid: Centro de Estudios Históricos, 1930. [572,000–250]

GEII (1260–1280) = Alfonso X. 1957. *General Estoria. Segunda parte*, 2 vols., ed. A. Solalinde, Ll. Kasten and V.R.B. Oelshläger. Madrid: CSIC (Consejo Superior de Investigaciones Científicas). [263,500–197]

Grimalte (~1486) = de Flores, Juan. 1971. *Grimalte y Gradissa*, ed. Pamela Waley. London: Tamesis. [25,500–10]

Cárcel (1492) = de San Pedro, Diego. 1972. *Cárcel de amor*, ed. Keith Whinnom. Madrid: Castalia. [25,500–5]

CRC (1482–1490) = Hernando del Pulgar, *Crónica de los Reyes Católicos*, 2 vols., ed. Juan de Mata Carriazo, Madrid: Espasa Calpe, 1943. [322,000–155]

Celestina (1499) = de Rojas, Fernando. 1987. *La Celestina*, ed. D.S. Severin, Madrid: Cátedra. [67,000–81]; Translated by Mack Hendricks Singleton, Madison: The University of Wisconsin Press, 1968.

Lozana (1528) = Delicado, Francisco. 1984. *La lozana andaluza*, ed. Bruno M. Damiani. Madrid: Castalia. www.cervantesvirtual.com. [64,000–50]

Quijote (1605–1616) = Miguel de Cervantes, *Don Quijote de la Mancha*. www.cervantesvirtual.com. [384,000–505]; Translated by Edith Grossman, *Don Quixote*, Harper Perennial, 2005.
Pepita (1870) = Valera, Juan. *Pepita Jiménez*. www.cervantesvirtual.com. [56,500–63]
Regenta (1870–1880) = Alas "Clarín", Leopoldo. *La Regenta*, vol I. www.cervantesvirtual.com. [141,000–258]
Perfecta (1876) = Pérez Galdós, Benito. *Doña Perfecta*. www.cervantesvirtual.com. [65,000–107]
Pazos (1886) = Pardo Bazán, Emilia. Los pazos de Ulloa. www.cervantesvirtual.com. [83,500–129]

References

Aijmer, Karin. 1997. *I think* – an English modal particle. In *Modality in Germanic languages: Historical and comparative perspectives,* T. Swan & O. Jansen Westvik (Eds), 1–47. Berlin: Mouton de Gruyter.

Boas, Hans C. 2003. *A constructional approach to resultatives.* Stanford CA: CSLI.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan. 1988. Morphology as lexical organization. In *Theoretical morphology: Approaches in modern linguistics*, M. Hammond & M. Noonan (Eds), 119–141. New York NY: Academic Press.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: CUP.

Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In *Handbook of historical linguistics,* B. Joseph & R. Janda (Eds), 602–623. Oxford: Blackwell.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 529–551.

Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2): 323–355.

Bybee, Joan & William Pagliuca. 1987. The evolution of future meaning. In *Papers from the 7th International Conference on Historical Linguistics*, A.G. Ramat, O. Carruba & G. Bernini (Eds), 109–122. Amsterdam: John Benjamins.

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world.* Chicago IL: The University of Chicago Press.

Camus Bergareche, Bruno. 2004. Perífrasis verbales y expresión del aspecto en español. In *El pretérito imperfecto*, L. García Fernández & B. Camus Bergareche (Eds), 511–572. Madrid: Gredos.

Carey, Kathleen. 1994. The grammaticalization of the Perfect in Old English: An account based on pragmatics and metaphor. In *Perspectives on grammaticalization, W. Pagliuca (Ed.), 103–117.* Amsterdam: John Benjamins.

Casenhiser, Devin & Goldberg, Adele E. 2005. Fast mapping of a phrasal form and meaning. *Developmental Science* 8(6): 500–508.

Company Company, Concepción. 2006. Zero in syntax, ten in pragmatics, or subjectification as syntactic cancellation. In *Subjectification: Various paths to subjectivity*, A. Athanasidou, C. Canakis & B. Cornillie (Eds), 375–398. Berlin: Mouton de Gruyter.

Dąbrowska, Ewa & Elena Lieven. 2005. Towards a lexically-specific grammar of children's question constructions. *Cognitive Linguistics* 16(3): 437–474.

Diessel, Holger & Michael Tomasello. 2001. The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics* 12(2): 97–141.

Dietrich, Wolf. 1983. *El aspecto verbal perifrástico en las lenguas románicas*. (Translated by Marcos Martínez Hernández.) Madrid: Gredos.

Erman, Britt & Ulla-Britt Kotsinas. 1993. Pragmaticalization: The case of *ba'* and *you know*. *Studier i modern sprakvetenskap 10: 76–92.*

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20: 29–62.

García Fernández, Luis (dir.), Ángeles Carrasco Gutiérrez, Bruno Camus Bergareche, María Martínez-Atienza & Maria Ángeles García García-Serrano. 2006. *Diccionario de perífrasis verbales*. Madrid: Gredos.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago IL: The University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: OUP.

Goldberg, Adele E., Devin Casenhier & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15: 289–316.

Goldberg, Adele E., Devin Casenhier & Nitya Sethuraman. 2005. The role of prediction in construction learning. *Journal of Child Language* 32(2): 407–426.

Goossens, Louis. 1992. *Cunnan*, *conne(n)*, *can*: The development of a radial category. In *Diachrony within synchrony: Language history and cognition*, G. Kellerman & M.D. Morrissey (Eds), 377–394. Frankfurt: Peter Lang.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39: 1041–1070.

Hopper, Paul J. 1991. On some principles of grammaticalization. In *Approaches to grammaticalization,* Vol.1, E. Closs Traugott & B. Heine (Eds), 17–35.Amsterdam: John Benjamins.

Hopper, Paul J. & Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge: CUP.

Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: OUP.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar,* Vol.1: *Theoretical perspectives*. Stanford CA: Stanford University Press.

Lieven, Elena, Julian Pine & Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of Child Language* 24(1): 187–219.

Lieven, Elena, Heike Behrens, Jennifer Speares & Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30(2): 333–370.

Marcos Marín, Francisco (dir.). 1992. COREC: Corpus de Referencia de la Lengua Española Contemporánea: Corpus Oral Peninsular. (www.lllf.uam.es/~fmarcos/informes/corpus/corpusix.html)

Myhill, John. 1988. The grammaticalization of auxiliaries: Spanish clitic climbing. *Berkeley Linguistics Society* 14: 352–63.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70: 491–538.

Olbertz, Hella. 1998. *Verbal periphrases in a functional grammar of Spanish*. Berlin: Mouton de Gruyter.

Scheibman, Joanne. 2000. *I dunno but* … a usage-based account of the phonological reduction of *don't* in conversation. *Journal of Pragmatics* 32: 105–24.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: OUP.

Spaulding, Robert K. 1926. History and syntax of the progressive constructions in Spanish. *University of California Publications in Modern Philology* 13: 229–284.

Squartini, Mario. 1998. *Verbal periphrases in Romance: Aspect, actionality, and grammaticalization*. Berlin: Mouton de Gruyter.

Tatlock, John S.P. & Kennedy, Arthur G. 1927. *A concordance to the complete works of Geoffrey Chaucer*. Washington DC: Carnegie Institute.

Thompson, Sandra A. 2002. 'Object complements' and conversation. *Studies in Language* 26(1): 125–164.

Thompson, Sandra A. & Anthony Mulac. 1991. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In *Approaches to grammaticalization, Vol.II,* E.C. Traugott & B. Heine (Eds), *313–329.* Amsterdam: John Benjamins.

Torres Cacoullos, Rena. 1999. Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. *Language Variation and Change* 11: 143–170.

Torres Cacoullos, Rena. 2000. *Grammaticization, synchronic variation, and language contact. A study of Spanish progressive -ndo constructions.* Amsterdam: John Benjamins.

Torres Cacoullos, Rena. 2001. From lexical to grammatical to social meaning. *Language in Society* 30: 443–478.

Torres Cacoullos, Rena. 2006. Relative frequency in the grammaticization of collocations: Nominal to concessive *a pesar de*. In *Selected proceedings of the 8th Hispanic Linguistics Symposium,* T.L. Face & C.A. Klee (Eds), 37–49. Somerville MA: Cascadilla Proceedings Project. (http://www.lingref.com/cpp/hls/8/index.html)

Torres Cacoullos, Rena & James A. Walker. 2009. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics 47.1*

Traugott, Elizabeth Closs. In press. Grammaticalization, constructions and the incremental development of language: Suggestions from the development of degree modifiers in English. In *Language evolution: Cognitive and cultural factors*, R. Eckardt & G. Jaeger (Eds), Berlin: Mouton De Gruyter.

Formulaic models and formulaicity in Classical and Modern Standard Arabic

Giuliano Lancioni
Department of Linguistics, Roma Tre University

Abstract

Classical Arabic [ClA], the historical antecedent of Modern Standard Arabic [MSA], is a language reconstructed from a selected, close textual corpus, clearly removed from everyday speech. The analysis focuses on the formulaic features ClA and MSA inherited from their models, which are consistently missing from spoken Arabic variants; these features range from text chunks to morphological and syntactic patterns (including redundant case affixes, and syntactically determined partial agreement). The general consequence of the hypothesis presented is that formulaicity in written languages can be strongly reinforced by the model of literary varieties, even long after the original textual constraints disappeared. The influence of MSA on modern spoken varieties shows the possibility that such formulaic features find their path through spoken languages.

1. Introduction*

Oral tradition studies, since Milman Parry's seminal work on the Homeric poems, have convincingly demonstrated that several poetic traditions are based to a large extent on oral composition and formulaic diction.

Despite some differences (on which see § 3.1. below), Parry's definition of the formula is couched in somewhat similar terms as contemporary ones within constructionist frameworks. However, oral formulas – with notable exceptions, e.g., Kuiper (2000) – are generally not recognized as phenomena relevant to linguistics, but rather as features at the stylistic level.

Whether or not this idea is right in the general case, formulas do matter in written languages based upon an oral-formulaic tradition. Classical Arabic is such a language: at least in some cases, formulas arisen in oral poetic contexts have very likely evolved into constructions which eventually found their way in the standard language and, through borrowing, even in spoken Arabic.

The present paper is organized as follows: after a short introduction on the role of formulas in the Classical Arabic poetical tradition and a sketch of the sociolinguistic situation of Arabic (§ 2), concurrent definitions of formulas in Parry-Lord theory and constructionist models are reviewed and shortly discussed (§ 3.1), together with the crucial question of the relationship between formulas and constructions (§ 3.2). The core part of the article (§ 4) illustrates some representative examples of Classical Arabic constructions for which I suggest a formulaic origin. In the conclusion, some perspectives for further study are shortly discussed (§ 5).

2. The oral-formulaic nature of the Classical Arabic texts

The technical definition of the formula as a basic building block in oral poetry is due to the groundwork laid by Milman Parry (see the writings gathered in Parry 1971), who first proposed that many of the puzzling and seemingly contradictory features of the Homeric poems could be explained if we consider them as the result of the written record of oral performances heavily based upon a stock of poetically specialized formulas. Fieldwork by Parry himself and his colleague

*The research upon which this work is based is partially funded by the Research Program of National Interest (PRIN) "Computer Analysis of the Hierarchical Structure of Arabic Lexicon: the Verbal System". I would like to thank Georges Bohas for reading a version of this paper and providing many helpful comments. Of course, all errors and all views expressed are my sole responsibility.

Albert Lord tested the theory, generally known as Oral Theory, in real contemporary situations (extensive accounts in Lord 1965; see also Ong 1982 for a wider discussion on orality vs literacy).[1]

Monroe (1972) first proposed that the oral-formulaic theory might be applied to the pre-Islamic oral poetic tradition. Zwettler's (1978) monograph has convincingly shown that this hypothesis is quite plausible and helps explain many difficult points in traditional accounts of the origin and transmission of pre-Islamic poetry. Paoli (2001) brought the support of computer scanning of several thousands lines of classical Arabic poetry, demonstrating what Zwettler had already shown for a more limited corpus, namely that these texts are formulaic, in Parry-Lord's sense, to a considerable degree. Even if the Monroe-Zwettler hypothesis is by no means uncontroversial in Arabic studies, I shall take for granted that Pre-Islamic poetry is formulaic.

Pre-Islamic poetry is, besides the Quran, the main source for Classical Arabic grammar and lexicon. Medieval Arab grammarians clearly recognized the partly artificial character of this variety, since they think the language has no native speaker as early as 650 AD, and regard Classical Arabic as a language reconstructed from a closed textual corpus, more or less removed from everyday speech.[2]

The Quran is a relatively short text: this reason, along with other considerations, led medieval Arab grammarians to make extensive use of Pre-Islamic texts, written in a variety that closely resembles the Quran's, as a reference material. As a consequence, a linguistic variety was largely built on a formulaic model. Since Classical Arabic is the ancestor of Modern Standard Arabic, to the point that many speaker do not consider them as two distinct varieties at all, the formulaic nature of Pre-Islamic texts percolated through to modern Arabic prose.

Deep contact with Classical and Modern Standard Arabic contributed to spread some of these, originally formulaic features as borrowings in spoken varieties.

3. Concurrent definitions of formulas

As it often happens with originally non-technical words, "formula" is far from having an unequivocal meaning. On the other hand, Wray & Perkins (2000: 3) list over 40 terms used in literature to describe formulaic sequences and formulaicity. In the case at hand, a crucial distinction is needed between the relatively similar, but far from

1. The presence of formulas is not per se an automatic index of orality: written texts may retain formulaic features for reasons of stylistic borrowings or whatever. See Lancioni (forthcoming) for discussion.

2. See Lancioni (2003) for a discussion of these and other related questions.

identical, definitions of formulas in Oral Tradition studies and in constructionist models (§ 3.1). Another important point review is the boundary between formulas and constructions, which I think is more blurred than usually assumed (§ 3.2).

3.1 What is a formula: Some definitions

A formula, in the Parry's own words, is "a group of phrases which have the same metrical value and which are enough alike in thought and word to leave no doubt that the poet who used them knew them not only as a single formula, but also as formulas of a certain type" (Parry 1971: 275). This definition, albeit slightly circular, gathers some of the structural features of formulas as originally detected by Parry in his pioneering studies on the Homeric poems:

1. formulas are open sequences of elements ("phrases") that are formally, semantically, metrically similar;
2. they are consciously employed in constant metrical positions.

Significantly, Parry does not require that formulas are verbatim identical in all their occurrences. To be a formula, it is sufficient that a "group of phrases" have "the same metrical value" and are "enough alike in thought and word".

Common practice in oral tradition studies has progressively tended to widen the concept of formulas, by relatively strengthening the prosodic and metric constraints while weakening collocational requirements. This kind of "open formulas" is in some respect closer to constructions than formulas as usually defined in constructionist paradigms. Recent literature on formulaic expressions, in fact, tends to endorse a stricter, more collocational definition of the formula, e.g., "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray & Perkins 2000). Formulas in oral tradition are considerably more flexible than Wray & Perkins' (2002) "transparent and flexible ones [i.e., formulas] containing slots for open class items", which do not go any further than such cases as *NP be-TENSE sorry to keep-TENSE you waiting* (Pawley & Syder 1983: 210).

Formulas in oral tradition studies are at least partially prefabricated in the sense that they are retrieved from memory as chunks: they are classes of prosodically, metrically and syntactically analogous forms which the poet-performer chooses according to his needs in order to quickly and appropriately fill a smaller or larger portion of a text. On the other hand, memorizing has a fundamental role in the formation of oral poets (and is specifically attested by numerous anecdotic tales on Pre-Islamic poets), and the needs of oral performance exert a

strong pressure on the oral poet,[3] who has to rely on ready-made material to be quickly recalled.

An interesting explanation of how the relatively open character of open formulas can be fit within the requirement that elements in the working memory be retrieved from memory as chunks is proposed by Kuiper (2000: 294): even if formulas have normal syntactic properties, they are inserted in working memory as finite state structures. This imposes a limitation on possible formulas, but does not prevent their intermingling with more articulated syntactic structures, since they can be reanalyzed as context-free or even more complex structures.

3.2 Formulas and constructions

Conventional wisdom about formulas seems to imply that they are of a clearly distinct kind than constructions:[4] constructions are open and related by a thick network of motivated relations (Goldberg 1995), while formulas are a stock of ready-made forms, directly retrieved in the speaker's working memory.

As we have seen, formulas in Parry's (and, more so, Lord's) terms are much closer to constructions than formulas in the general linguistic sense are.[5] What distinguishes them from the latter is their belonging to a fixed, albeit evolving, stock and their limitation to specific prosodic and metric contexts: two restrictions which quickly fade away when formulas are severed from their original poetic context.

My hypothesis is that formulas naturally evolved into constructions when the grammar of a written language was built upon a corpus of formulaic, oral poetic texts. What was an open formula became, when detached from its original context, a construction which could later develop secondary meanings and give rise to other, related constructions, within the limits imposed by the unavoidable conservatism of a relatively artificial, learned language.

---

3. In Albert Lord's (1960: 65) words, "The singer's mode of composition is dictated by the demands of performance at high speed".

4. The difference between constructions and formulas is felt so strong that Wray (2002) advocates the need for entirely different learning approaches for each in second language teaching. In this connection, Granger (1998: 157) even says that "there does not seem to be a direct line from prefabs to creative language" in language learning.

5. Moreover, oral tradition scholars do not usually regard as formulas prefabricated items in ordinary speech. The rate of "formulaic density" (which rarely climbs much over 30%; see Lord 1986: 478–481) is often much lower than analogous counts in constructionalist framework (where rates as high as 80% have been reported).

In this respect, Bybee's (2002, 2003) concept of "grammaticalization as automatization" is very useful, albeit in a slightly modified sense. Bybee's original use of the concept is about the gradual fixing of constructions out of statistical regularities. In the case of Classical Arabic, constructions have been recognized, classified and have become gradually productive because of their frequency in the reference corpus.

The general schema can be summarized as in (1):

(1)	Coinage	→ formula	→ construction	→ expansion
	a poet/performer creates a sequence by invention or variation	the sequence finds its way into the poetic stock	the standardization of the poetic corpora transforms the formula into a construction	the constructions may become productive and give rise to other constructions

Let us briefly comment upon this schema. First, a formula was born through coinage. This is a very important, even if often neglected aspect, of the production of an oral-formulaic language: despite their persistence through time – to the extent that they may preserve even very archaic features, long disappeared from everyday language – formulas are not at all invariable. It is the slow evolution of the stock of poetic formulas through time which gives formulaic language their most striking aspect: the coexistence of diachronically and diatopically incoherent features – archaisms together with much more recent forms, even at the level of flexional morphology (e.g., the different forms of Genitive in Homer), expressions belonging to different dialectal varieties, so-called solecisms, and so on.

In unpublished papers, Fillmore (1997) and Kay (2002) have argued against the confusion among coinages and constructions. While in general Fillmore and Kay's skepticism is to be accepted, a gradual, extremely selective passage from coinages to formulas is a necessary assumption for Oral Theory – if the very idea of oral-formulaic poetry is to be supported. In some cases at least, the boundary between coining and construction needs not to be too rigid. For entirely independent reasons, Leino (2005: 6), in his study on the naming patterns of Finnish lakes, states that "their [Fillmore's (1997) and Kay's (2002)] distinction between productive constructions and unproductive patterns of coining seems to be too strict for my present needs: the phenomenon I am trying to describe is neither systematically productive nor unproductive, but somewhere between these extremes."

The passage from the second to the third box in (1) is even harder to fit in current assumption within constructionist models. Formulas are in general regarded as rigid structures, and one can often read that "language is more formulaic than creative" (as in Wray 2002) or "formulaicity contrasts with productivity" (Wray &

Perkins 2000). On the contrary, formulas in oral poetry are by definition created, even if the share of novel formulas is very little in any oral poet's formulaic stock.

4. Some examples

In the next sections we shall review some of the Classical Arabic constructions for which I hypothesize a formulaic origin. These examples are to be seen as a sampling, rather than an exhaustive classification, and have been chosen in order to show cases of increasing complexity and structural significance: first the relatively marginal phenomenon of dual agreement marking in verbs and adjective (§ 4.1), then the singular, hard-to-explain coincidence of form and traditional denomination of some case endings in nouns and mood vowels in verbs (§ 4.2). The next two sections examine two of the syntactic phenomena which are harder to explain within both the traditional and the contemporary linguistic framework, namely partial agreement in VS contexts (§ 4.3) and Genitive case on nominal governed by the conjunction *wa-* 'and' (§ 4.4). The final section is devoted to the single most important feature which distinguishes Classical and Modern Standard Arabic from any spoken variety, namely the presence of a system of case endings on nominals, which I think can get a better explanation from a formulaic origin than from what has been done up till now in the literature (§ 4.5).

4.1 Dual agreement

Classical Arabic has a fully functional dual number, which requires dual agreement in verbs, adjectives and pronouns. The presence of dual number in nominals varies wildly in Arabic dialects, from virtually dualless varieties (e.g., Moroccan Arabic) to varieties where dual is regular and productive (e.g., Syrian Arabic); no Arabic dialect, however, shows dual agreement in verbs, adjectives and pronouns. Dual nominals mostly agree with plural verbs, adjectives and pronouns instead.

Traditional accounts interpret this phenomenon in terms of "loss" of dual agreement. However, some features clearly speak against this historical account. First, comparing within Semitic languages shows that only nominal dual could be regarded as possibly Proto-Semitic, while full dual agreement in Arabic might be properly regarded as an extension of this agreement mechanism.[6]

Second, full agreement is not a regular feature, nor in Classical Arabic nor in Arabic dialects. For instance, plural nonhuman nominals in Classical Arabic tend

6. Cf. Ferguson (1959: 620, fn. 8); Blanc (1970).

to agree, in MSA almost always do so, with feminine singular verbs, adjectives and pronouns.[7] Or, verbs preceding their nominal subjects always remain singular (this is the well-known phenomenon of partial agreement).

(2) a. bēt-ēn kbār (Syrian Arabic: Ferguson 1959: 621)
house-DU large/PL
'two large house'

b. bayt-āni kabīr-āni (Classical Arabic)
house-DU;NOM large-DU;NOM

An alternative account I would like to propose is that dual agreement arises from a formulaic extension of dual markers on nouns. This extension could very well be due to prosodic, as well as to stylistic reasons: the constant repetition of the *-āni -ayni* endings produces both regular –∪ prosodic patterns and internal rhyming effects.[8]

In this specific case, poetry is not the most likely source of the pattern, because of the immediate rhyming effect which contrast with the verse-final rhyme, constant through the whole poem, which characterizes pre-Islamic and Classical Arabic metrics.

The root for this phenomenon is rather to be found in a distinct, albeit closely related, tradition, namely *sa'*. By this term, which will later be reinterpreted and understood in the sense of "rhythmic and rhyming prose", was designed in pre-Islamic age a kind of oracular style, typical of soothsayers, which was characterized by internal rhythmic and rhyme effects. We have very scanty remnants, mostly unsure, of *sa'*, but the style is generally regarded as very close to what can be found in the older chapters of Quran.

Here is an example from the Surah of the Cave (Qur. 18.82, trans. Pickthall):

Wa-ammā 'l-ǧidār-u fa-kāna li-ġulām-ayni yatīm-ayni
And-TOPIC DET-wall-NOM COMMENT-was to-boy-DU;GEN orphan-DU;GEN
fī 'l-madīnat-i
in DET-city-GEN
"And as for the wall, it belonged to two orphan boys in the city"

Without deepening any further into a complex, highly controversial issue, we can reasonably hypothesize that this Quranic usages reflects a *sa'*-like stylistic

7. Several agreement patterns are possible in this context in Arabic dialects, e.g., plural feminine. See Blanc (1970) for details.

8. The best, most complete account of Classical Arabic metrics is Bohas & Paoli (1997).

pattern, which was later grammaticalized in the process of standardization of the Quranic text.[9]

In this case, we may imagine the following development:

1. dual endings are added to adjectives as a stylistic variation to meet the needs of *sa*ʿ; this is very likely, because nouns and adjectives are formally indistinguishable in Arabic, and adjectives can be treated as nouns without any specific device;
2. coinages with adjectival dual gradually make their way into the stock of formulaic expressions;
3. dual endings are extended by stylistic analogy to imperfect verbal form, which are, and have always been recognized as, relatively closer to nouns (Arab grammarians noted the identity of modal/case vowels of, respectively, nominative/indicative and accusative/subjunctive, a feature which might be thought as arising from formulaic usage as well, since it is not attested in any Arabic dialect);

Table 1. Possible origin of verb dual

	nominative/indicative	**accusative/subjunctive**	**dual**
noun	*kitāb-u*	*kitāb-a*	*kitāb-āni*
verb	*yaktub-u*	*yaktub-a*	***yaktub-āni***

4. the stylistic variation becomes more and more widespread because of its prosodic features (regular –∪ pattern) and becomes an alternative construction;
5. grammarians find the dual agreement construction as fitting their system and transform it in the standard agreement "rule".

This account, in my opinion, explains reasonably well some striking features of the employ of dual agreement on verbs and adjectives, namely that it is foreign to spoken Arabic – and is felt by native speakers as a mark of the written language to a considerable degree – yet it is exceedingly rare in Classical poetry: its origin from a limited, *sa*ʿ-oriented portion of the Quranic text is able to account for both features, despite their seeming contrast.

4.2 Imperfect–noun analogy

Arab grammarians draft a parallelism between case endings of nominals and mode endings of imperfect verbs. In particular, they call in the same way, respec-

9. See Bohas (in preparation) for an analysis of Surah Raḥmàn of the Quran which detects an *–ān* rhyme pattern that is consistent with my hypothesis.

tively, nominative and indicative (*raf'*), which both end prototypically in *–u*, and accusative and subjunctive (*nab*), which both end in *–a*.[10]

Table 2. Analogical forms between nouns and imperfect verbs

	nominative/indicative	**accusative/subjunctive**
noun	*kitāb-u*	*kitāb-a*
verb	*yaktub-u*	*yaktub-a*

The parallelisms is further strengthened by the fact that imperfect tense is called *muḍāri'*, literally 'resembling', because it should make the verb resemble a noun (which is rationalized by its use in nominalizations).

Since modal oppositions are not realized this way in Arabic dialects (some dialects mark modal oppositions by preverbs or adverbs instead), a formulaic origin for the imperfect verb endings is not unlikely. In particular, a prosodic constraint which tended to prefer open over close syllables, together with the necessity to avoid consonant clusters at word boundaries, might have prompted the addition of final vowels to imperfect verbal forms (which by themselves in most cases lacked an independent one, unlike perfect forms).

Let us summarize:

1. a final vowel is required by prosodic contexts; an epenthetic vowel is more or less randomly added, perhaps according to vocalic harmony considerations;
2. *-a* and *-u* tend to prevail, perhaps for similarity with noun case endings (see below);
3. a specialization of vowels for modal purposes is established by grammarians, which transform a constructional *cliché* into a "rule".

It is important to note that the functional yield of modal oppositions in verbs, as the yield of case oppositions in nouns, is very low, being most often accompanied by selection or conjunctions and/or negations. Moreover, modal distinctions are often partially or entirely blurred (notably in most dual and plural forms).

10. Carter (1981), in his commented translation of an introductory medieval Arab grammarian by al-Širbīnī, tries to maintain the parallelism between nominative/indicative and accusative/subjunctive by translating them with, respectively, 'independent' and 'dependent'.

4.3 Partial agreement in VS contexts

A very striking feature of Classical and Modern Standard Arabic syntax is the lack of verb number agreement in syntactical contexts where the verb precedes a nonpronominal expressed subject. This feature, which is extraordinary difficult to explain within syntactic models (see Mohammed 1990, Lancioni 1996 for competing syntactical accounts), cannot be found in other Semitic languages or in Arabic dialects. An important condition for this form of agreement is the relative position of verb and subject: if the former precedes the latter partial agreement obtains, while in case of reverse order the verb fully agrees with its subject.[11]

Since Classical Arabic is a predominantly VS language, a formulaic origin of this partial agreement is more than likely. Singular verbs at the beginning of a verse, perhaps after a conjunction, are extremely frequent; a *cliché* which always uses a singular verb, independently from the number of the following subject, has much wider formulaic potentialities than a fully agreeing verb.

Differences between spoken and Classical Arabic are exemplified in (3):

(3) a. ʾažū=na ḍyūf min ʾIṭālya (Palestinian Arabic, Durand 1996: 155)
came.3MP=1P host.P from Italy
'Hosts came and visited us from Italy'

b. ǧāʾa=nā ḍuyūf-u-n min Īṭāliyā
came.3MS=1P host\PL-NOM-INDF

That a formulaic power is really working here is hinted at by another interesting agreement, or rather disagreement, phenomenon recorded by Classical Arab grammarian: namely, the possibility to have a verb in the third masculine singular even with a feminine subject (that is, a verb lacking not only number, but also gender agreement), provided the subject is separated by the verb by at least an intervening word (e.g., an adverb).

Full agreement is recognized as an alternative by Arab grammarians themselves, since they record a fully agreeing construction which they dub with the catchword *akalūnī l-barāġīṯu* 'the flies ate me', with a plural verb agreeing with a following plural subject. Albeit a minority variant, this construction is regarded as grammatical.

11. As usual in literary traditions, the picture of agreement in Classical Arabic is more complex than this sketchy description would suggest: see Bohas (2007) for details.

Some typical cases where such a formulaic structure is attested in Pre-Islamic poetry have been listed by Zwettler (1978: 52). For instance:

(4)	Metric schema	∪ – ⏑	∪ – ⏑	– ∪ – ⏑	∪ – ∪ –
	Arabic formula	*wa-/fa-* + *zalla*.PRF.3ms 3ms-*zalla*.EPFV-IND	subject_NP/PL	verb.PL	PP-object_NP
	Paraphrase	'and continued/continues'	'Xs'	'(they) verb'	'(with/to) Z'
	Example	*wa-alla*	*ṣiāb-ī*	*yaštawūna*	*bi-niʿmatin*

'And my friends continued to roast with pleasure …'

The origin of this formula can be found in the iteration of coordinated singular verbs at the beginning of subsequent verses which leads to the rationalization of a VS constraint.[12]

Let us sum up what might be happened through this process:

1. verse-initial *clichés* tend to use a singular verb (or even a masculine singular verb) in order to be applicable to wider contexts;
2. gradually the *cliché*, because of the higher frequency of verse-initial contexts, spreads across other, non-verse-initial, contexts, becoming kind of a construction;
3. singular verb construction gradually overcomes full agreement verb construction, which becomes a marked construction;
4. since this construction is originally, and predominantly, verb-initial, Arab grammarians establish the "rule" which ties the choice between partial and full agreement to the fact that the verb precedes or follows its subject.

This possible formulaic origin for the partial agreement construction has the advantage to give a textual account to a structure which is doubtless hard to explain – and which has been given no completely satisfying account both in traditional grammar and in contemporary linguistics – within the general framework of Arabic syntax.

4.4 *Wa* 'and' = *rubba* 'many'

The usage of the conjunction *wa-* 'and', with a subsequent genitive case noun, in the meaning of 'many' is characteristic of Classical Arabic, and is not attested in

12. The preference for a singular verb in this case is not explained by the metrical context (a plural verb would fit the meter as well), but rather by the practice of repeating several verses with a similar initial pattern. I would like to thank Georges Bohas for pointing my attention to this question.

Arabic dialects or other Semitic languages. This constructions looked puzzling to Middle Ages Arab grammarians, since according to their principles of syntactic government only an overt of covert preposition can govern genitive case, and a lot of discussion and ingenuity were needed to find viable solutions to explain the existence of this structure.

Here is the general schema of this formula (again from Zwettler 1978: 52):

(5)	Metric schema	∪ – –	∪ – –	– ∪
	Arabic formula	*wa-* + noun.SG-GEN-INDF	verb ± pronoun	noun-NOM/ACC
	Paraphrase	'and an X' →'many Xs'	'(that) verb'	'X/Y'
	Example	*wa-far'-i-n*	*yazīnu*	*l-matna*
		'And a perfect head of hair which [when loosened] adorns her back' → 'Many heads of hair …'		

In this case, the formulaic origin of the construction is particularly clear: the genitive case is clearly not governed by *wa-*, for both the general syntactic patterns in Arabic and the lack of the construction elsewhere than in Classical Arabic.[13]

What can be assumed quite reasonably in this case is that the formula arose out of the enumeration of a series of genitive complements at the beginning of subsequent verses governed by the same preposition many verbs in Classical Arabic govern a prepositional phrase, which in the Arab grammatical tradition gave birth to the category of "transitive through a preposition". The frequency and metrical expediency of such a structure gradually transformed it into a *cliché* detached from the governing verb. When the formula was rationalized by grammarians it was converted, not without pain, as we have seen into a new construction.

As happened elsewhere, the legitimating of this structure within the grammatical standard caused its expansion in textual types other than poetry, e.g., ordinary prose texts.

4.5 A provocative solution to an enigma: the birth of case endings

One of the most striking phenomena which set apart Classical Arabic and Arabic dialects is the lack of case-endings in the latter. Classical Arabic has a three-case system which in most cases (singular and broken plural nominals) encodes case with a single-vowel ending, *-u* for Nominative case, *-a* for Accusative case, and *-i* for Genitive case (usually called Oblique case in the Orientalist traditional terminology).

13. The first gloss reports the original meaning of *wa-* as a copulative conjunction in a clearly enumeration-like pattern.

Minor classes of singular and broken plural nominals (so-called dyptote nominals when they are indefinite), along with dual and sound plural nominals, reduce to a two-ending system, where Accusative and Genitive case collapse, while the difference among the two distinct ending reposes always on a vocalic alternation, perhaps with a final added element (*-ni* in the dual and *-na* in the masculine sound plural) which falls in some contexts (in particular, in the possessive construction known as construct state).

This system cannot be found in Arabic dialects, where singular, broken plural and feminine sound plural nominals have a zero ending, while dual and masculine sound plural nominals have endings which correspond to the accusative/genitive forms (*-ayn* for the dual and *-īn* for the plural, respectively).

Table 3. Nominal flexion in Classical and spoken Arabic

	Tryptote nominals	**Dyptote nominals**	**Dual**	**Masculine Sound Plural**	**Feminine Sound Plural**
Nom.	*-u*	*-u*	*-ā(ni)*	*-ū(na)*	*-āt-u*
ACC.	*-a*	*-a*	*-ay(ni)*	*-ī(na)*	*-āt-i*
Gen.	*-i*				
Arabic dialects	*-Ø*		*-ayn*	*-īn*	*-āt*

Semitic studies have traditionally thought this phenomenon in term of "loss" of case-endings in the spoken varities, since a system similar to the Classical Arabic one can be found in the oldest-attested Semitic language, namely Akkadic. In this respect, Arabic would be a very conservative language, despite being attested relatively late, since most Semitic languages have no case system at all.

This view, which considers the case-system in Classical Arabic as a very archaic feature, is partially defied by epigraphic data, which show that already in 3rd century BC, that is roughly nine hundred years before the oldest attested Classical Arabic poems, Arabic did not have a fully functional case system. In Nabatean inscriptions from that period, in fact, Arabic proper names are written with seemingly random vocalic endings, without any apparent case system.[14]

Moreover, Corriente (1971) has convincingly shown that case-endings in Classical Arabic texts (including Classical poetry) have a "functional yield" close to zero: that is, in almost every context they are entirely redundant, and do not

14. The relevant data from Nabatean inscriptions are reviewed and extensively discussed by by Diem (1973).

significantly contribute to the identification of the right syntactic context.[15] According to his view, the case forms are not well integrated into the morphology; and since they are marked by a "lack of allomorphy" (see Table 3 above) which contrasts with the general fusional character of Arabic, one can reasonably assume that they are not originally cases, but epenthetic vowels.

The idea that cases are primarily epenthetic in nature is not without anticipations in the Arab grammarians' reflections on their own language. Already the grammarian Quṭrub is said to have claimed that cases in Classical Arabic serve primarily for prosodic reasons. Quṭrub is certainly to be regarded as a "dissenting grammarian", as Versteegh (1981) dubs him, but it is in my opinion highly significant that doubts on the functional nature of ClA case-endings were already present in the grammatical tradition. Guillaume (1998), on the other hand, shows clearly that the debate on the function of cases and the relationship between their form and function flourished among classical grammarians.

Also Owens (1998) questions the traditional account, even if from a different point of view, and hypothesizes that the caseless system in modern Arabic dialect goes "back to a caseless version of proto-Semitic". In his account, already in Proto-Semitic a split had taken place between two varieties, one with a case system and a caseless one.

According to him, however, the original distinction in case varieties was between the bare nominal stem (with zero suffix) and a case-marked form in *–a*, which is the antecedent of Arabic Accusative.[16] This alternance would be reflected in the opposition between a pausal (vowelless) and a nonpausal form (ending in a vowels) which is customary in Classical Arabic orthoepy.

In Owens' (1998: 71) own words: "It is precisely in this lack of symmetry that one might search for the origins of the Arabic case system (proceeding on the assumption that case in Semitic, where it exists, is innovative). This pausal

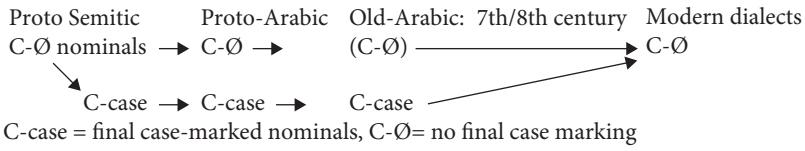
---

15. Corriente's (1971) argumentation is mainly based on a "commutation test" which shows that the proper parsing of Classical Arabic sentences is in most case independent of the information given by case-endings. Though this analysis have been sharply debated, I maintain Corriente is basically right in this respect. The very fact that most case-endings are simply ignored by standard writing – unless short vowels are marked, which happens in Quran and, less regularly, poetry only, and which did not happen in the oldest remnants of Arabic script – is a clear hint of the nonessential functional role of these markers.

16. The special status of the suffix *–a* is shown by its, sometimes fossilized, presence in Semitic languages which do not have a regular case system (e.g., Hebrew, where it can be found in some adverbial formations), and even in Arabic dialects – where it can however be interpreted as a borrowing from the Classical language.

alteration may represent an older state of affairs where an *–a(a)* suffix (as seen above, representing the unmarked case in Arabic) was opposed to a bare nominal stem (Ø). The nominative and genitive vowels may then have developed out of epenthetic vowels which were inserted in particular contexts."

Table 4. Development of case in Arabic (from Owens 1998: 224)



C-case = final case-marked nominals, C-Ø= no final case marking

In my opinion, the epenthetic nature of (at least some) case endings in Arabic is the crux of the question. In particular, the Nominative and Genitive case endings – *-i* and *–u* respectively – are very likely candidates to an epenthetic origin, even because the vowels *i* and *u* get confused in most dialect (e.g., in most positions they become *ə* in Syrian Arabic), even in those cases where a distinction among short vowels is preserved.

The origin of this epenthetic phenomenon can be traced to a restructuration of the syllable structure.

Though Semitic languages share a common morphological structure to a remarkable degree, they show very different models of syllable organization. Some languages have phonological system which are pivoted around a set of distinctly articulated vowels, both short and long, which tend to distribute in a limited array of syllable varieties (mostly CV and CVC), while other languages show a wider variety of syllable types, with a limited role played by vowels, especially by long ones, and a tendency to less distinct short vowels.

Classical Arabic and Arabic dialects clearly stand on opposite sides in this regard. Classical Arabic has a very symmetric vocalic system, with three short vowels (*a*, *i*, *u*) and their long counterparts. Since consonantal clusters are not admitted after word boundaries, the syllabic types reduce to two, namely CV and CVC. Biconsonantal clusters are possible only at the boundary of two syllables, while clusters of three or more consonants are not allowed. In some cases, prosthetic vowels are required by the orthoepic norms of Classical Arabic in order to avoid "prohibited" sequences.

Arabic dialects, on the other hand, all show a reduction of the role of vowels, albeit to different degrees. Western dialects often have a single short, schwa-like vowel, while Eastern dialect generally have a fuller inventory of short vowels. In any case, even variants which do not blur the distinctions among short vowels show a considerably wider syllabic inventory than Classical Arabic one. In the

following table some of the differences in possible syllable types among Classical and Syrian (Damascus) Arabic, the latter being an Eastern, rather "vocalic" variety, are shown:

Table 5. Examples of different syllable structures in Classical and Syrian (Damascus) Arabic

Classical Arabic		**Syrian Arabic**		
fahimtu	CV.CVC.CV	fhəmt	CCVCC	'I understood'
(bi-)taktubī	(CV.)CVC.CV.CV	btəktbi	CCVC.CCV	'you (fem.) write'
ṭāliba(tun)	CV̄.CV.CV(.CVC)	ṭālbe	CV̄C.CV	'student (fem.)'

Summing up, let us try to sketch a formulaic theory of the birth of case endings in seven steps:

1. Arabic used in poetry tends to alter its syllabic structure in order to meet the needs of a metrically regular diction.
2. Since the syllabic types in poetry reduce to CV and CVC, with a clear preponderance of the former type, verses show a tendency to have CV final syllables to comply with the needs of the rhyme system.
3. Epenthetic vowels are first introduced in final position in a rather random way, perhaps extending adverbial marker *-a* under the influence of constraints of vocalic harmony (see Quran). At first, this can arise through individual coinage.
4. Rhyme exigencies force final vowels to be constant across a poem. Coinage gradually makes its way into the thesaurus of poetical formulas.
5. Generalization across formulaic constructions gradually transforms prosodically motivated final vowels into constructional *clichés*.
6. The work of the Arabic grammatical tradition gradually rationalizes *clichés* and transforms them into "grammar rules", or regular constructions (but the "dissenting grammarian" maintains that cases in Classical Arabic serve primarily for prosodic reasons).[17]

17. A difficulty to my hypothesis I am aware of is that similarity of case-endings among Classical Arabic and other Semitic languages with morphological case systems (Akkadian, Ugaritic) would arise casually. A possible solution is to posit some kind of optimality condition along the lines already shown by Ibn Ǧinnī within the Arabic grammatical tradition, namely that the match between vowels and cases is determined by economy reasons according to a constraint of 'heaviness'. See the discussion in Bohas and Guillaume (1984), Bohas et al. (1990: 73–93).

7. Classical works on philosophy of language (since the 10th century) begin to try to match form and function of cases after the fact. Constructions are fully inserted into the machinery of Arabic grammar.

This account, even if it is far from many mainstream views of the case system in Arabic, is in my opinion able to explain in a more plausible way many of the aspects in the development of the case endings. First of all, the notion that most case endings originated from epenthetic vowels eliminates the need to postulate a historical continuity from Akkadic to Arabic, which would have proceeded without any trace in other varieties for millennia: in my account, we should speak of an independent "birth" of case endings, rather than of the development of older forms. Moreover, Classical and spoken Arabic would derive from a common stem, without any need to hypothesize a "loss" of case-endings for dialects.[18]

5. Perspectives

The examples of possible transformation of poetical formulas into constructions have been hitherto exclusively illustrated by examples from Arabic. However, the process which has been shown can be of more general interest, and might reasonably be applied to a wider spectrum of linguistic phenomena in different languages.

With all its specificities, Arabic is no special case in this respect. Other written languages have been standardized on the basis of textual corpora: the importance of the Vedic texts for Sanskrit or the Bible for Hebrew can hardly be overestimated. In these cases too, it is almost unthinkable that at least some idiosyncratic phenomena in corpora could not find their way in the standardized, written language – and again in spoken languages by borrowings (e.g., Modern Hebrew).

It is the general process of intermingling and reciprocal influences between literary formulaic languages and ordinary language which, in my opinion, deserves further research and investigation. Such a study is likely to shed a fresh light on several important, yet neglected, aspects of the nature of formulaic languages and their relations with distinct textual domains.

---

18. As Zwettler (1978) points out, examples of Arabic oral poetry with seemingly random ending vowels have been reported as late as the beginning of the 20th century. Even in this case, if we accept the idea that oral poetry uses final vowels for mainly prosodic reasons, these phenomena would rather be interpreted as cases of continuity with an older poetic tradition, which has not resented of the normative intervention of grammarians.

References

Blanc, Haim. 1970. Dual and pseudo-dual in the Arabic dialects. *Language* 46: 42–57.

Bohas, Georges & Jean-Patrick Guillaume. 1984. *Etude des théories des grammariens arabes*, Vol. 1, *Morphologie et phonologie*. Damas: Institut français.

Bohas, Georges, Jean-Patrick Guillaume & Djamel Eddine Koulughli. 1990. *The Arabic linguistic tradition*. London: Routledge.

Bohas, Georges. 2007. Sur une conception restrictive de la langue arabe. *Langues et Littératures du Monde Arabe* 6: 35–51.

Bohas, Georges. In preparation. *La mesure de la sourate Al-Raḥmân*.

Bohas, Georges & Bruno Paoli. 1997. *Aspects formels de la poésie Arabe*, Toulouse: AMAM.

Bybee, Joan L. 2002. Main clauses are innovative, subordinate clauses are conservative. In *Complex sentences in grammar and discourse*, J.L. Bybee & M. Noonan (Eds), 1–17. Amsterdam: John Benjamins.

Bybee, Joan L. 2003. Cognitive Processes in Grammaticalization. In *The New Psychology of Language* II, M. Tomasello (Ed.), 145–167. Mahwah NJ: Lawrence Erlbaum Associates.

Carter, Michael G. 1981. *Arab Linguistics: An introductory classical text with translation and notes*. Amsterdam: John Benjamins.

Corriente, Federico. 1971. On the functional yield of some synthetic devices in Arabic and Semitic morphology. *Jewish Quarterly Review* 62: 20–50.

Diem, Werner. 1973. Die nabatäische Inschriften und die Frage der Kasusflexion im Altarabischen. *Zeitschrift des Deutschen Morgenländischen Gesellschaft* 123: 227–237.

Durand, Olivier. 1996. *Grammatica di arabo palestinese. Il dialetto di Gerusalemme*. Roma: Università degli Studi 'La Sapienza'.

Ferguson, Charles A. 1959. The Arabic koine, *Language* 35: 616–630.

Fillmore, Charles J. 1997. *Lecture on idiomaticity*. Notes available at: http://www.icsi.berkeley.edu/kay/bcg/lec02.html.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, analysis and applications*, A. Cowie (Ed.), 145–160. Oxford: OUP.

Guillaume, Jean-Patrick. 1998. Les discussions des grammairiens arabes à propos du sens des marques d'iʿrāb. *Histoire Épistémologie Langage* 20: 43–62.

Kay, Paul. 2002. Patterns of coining. Paper presented in Second International Conference on Construction Grammar. Electronic version: http://www.icsi.berkeley.edu/kay/coining.pdf.

Kuiper, Koenraad. 2000. On the linguistic properties of formulaic speech. *Oral Tradition* 15: 279–305.

Lancioni, Giuliano. 1996. Arabic and Celtic sentence structure: The generalized expletive hypothesis. In *Studies in Afroasiatic Languages*, J. Lecarme, J. Loewenstamm & U. Shlonsky (Eds), 135–158. The Hague: HAG.

Lancioni, Giuliano. 2003. Oralità e scrittura, apporti esterni, concettualizzazioni innovative. In *Lo spazio letterario del Medioevo,* 3: *Le culture circostanti, II. La cultura arabo-islamica*, M. Capaldo, F. Cardini, G. Cavallo & B. Scarcia Amoretti (Eds), 233–258. Roma: Salerno.

Lancioni, Giuliano. Forthcoming. Variants, links, and quotations: Classical Arabic texts as hypertexts. In *Festschrift for Biancamaria Scarcia Amoretti*, D. Bredi, L. Capezzone, L. Rostagno & W. Dahmash (Eds).

Leino, Antti. 2005. In search of naming patterns: A survey of Finnish lake names. In *Denominando il mondo. Dal nome comune al nome proprio/Naming the world. From common nouns to proper*

names [Quaderni Internazionali di RIOn 1], D. Brozovic-Roncevic & E. Caffarelli (Eds), 355–367. Roma: Societa Editrice Romana.

Lord, Albert. 1960. *The singer of tales.* New York NY: Athenaeum.

Lord, Albert. 1986. Perspectives on recent work on the oral traditional formula. *Oral Tradition* 1: 467–503.

Mohammad, Mohammad A. 1990. The problem of subject-verb agreement in Arabic: Towards a solution, In *Perspectives on Arabic Linguistics* I, M. Eid (Ed.), 95–126. Amsterdam: John Benjamins.

Monroe, James T. 1972. Oral composition in pre-islamic poetry. *Journal of Arabic Literature* 3: 1–53.

Ong, Walter J. 1982. *Orality and literacy: The technologizing of the word*. London: Methuen.

Owens, Jonathan. 1998. Cases and proto-Arabic. *Bulletin of the School of Oriental and African Languages* 61: 51–73, 215–227.

Paoli, Bruno. 2001. Meters and formulas: The case of ancient Arabic poetry. In *Linguistic Approaches to Poetry* [Belgian Journal of Linguistics 15], C. Michaux & M. Dominicy (Eds), 113–136. Amsterdam: John Benjamins.

Parry, Adam (Ed.), 1971. *The making of the Homeric verse: The collected papers of Milman Parry*. Oxford: Clarendon.

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: native-like selection and nativelike fluency. In *Language and communication*, J.C. Richards & R.W. Schmidt (Eds), 191–226. London: Longman.

Versteegh, Kees. 1981. A dissenting grammarian: Qutrub on declension. *Historiographia Linguistica* 8: 403–429.

Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20: 1–28.

Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21: 463–489.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.

Zwettler, Michael. 1978. *The oral tradition of classical Arabic poetry*, Columbus OH: Ohio State University Press.

A corpus study of lexicalized formulaic sequences with preposition + *hand*

Hans Lindquist
Växjö University, Sweden

Abstract

In this study all n-grams including the lemma HAND were automatically retrieved from the British National Corpus and then formulaic sequences were manually distinguished. A special focus was given to *at hand*, *in hand*, *on hand* and *to hand* which occurred with reference to humans, animals, concrete inanimate items and abstract items. The abstract meanings were metonymic and metaphorical mappings of literal body part meanings onto more abstract meanings.

The data indicate that these sequences are the result of a lexicalization process in which they are developing towards univerbation. They seem to be processed holistically without regard to the meaning of the preposition, which is supported by the fact that they are now occurring written solidly in informal registers.

1. Introduction

As pointed out by e.g., Wray (2002: 7), the tendency by speakers to use groups of words together as multi-word units or phrases is a long-recognized phenomenon

which has been acknowledged by de Saussure, Jespersen, Bloomfield and many others (cf. Wray 2002: 7 for references). Wray tries to find the rationale behind the human inclination to use these "words and strings of regular as well as irregular construction" (2002: 279), and suggests that the explanation lies in the human learning and storing process on the one hand and the retrieval process on the other (2002: 100–102).

Corpus linguistic approaches to multi-word units include Sinclair (1991, 1999, 2003), Moon (1998), Hunston & Francis (2000) and Stubbs (2001, 2002). Much of this work was inspired by Palmer's (1933) and Firth's (1957) work on collocation. These studies have different aims, but they all try to understand how lexical items go together in ways which are decided by semantics, not by syntax (or not solely by syntax). The different approaches have given rise to a plethora of terms. In this paper I will mainly use the term 'n-gram' for recurring strings (with or without linguistic structure) that can be found in corpora and 'formulaic sequence', or 'phrase' for meaningful, linguistically structured recurring strings of words.

Most of the theoretical advances mentioned have been facilitated by the development of computers and computer corpora which have made it possible to replace introspection and dictionary data with authentic language in context. Even scholars with a background in the generative paradigm now recognize the value of corpus work, e.g., Wasow (2002: 163): "[…] given the abundance of usage data at hand, plus the increasingly sophisticated search tools available, there is no good excuse for failing to test theoretical work against corpora." The present investigation, however, goes further than this in that it also uses corpus searches as a discovery procedure, thus being to some extent corpus-driven (for discussions of this methodology, see Tognini-Bonelli 2001; Mair 2006: 33–35; Lindquist 2007). In this way, it is possible to extract not only well-known formulaic expressions and idiomatic phrases (like the ones studied by e.g., Moon 1998) but also frequently recurrent combinations of words which are syntactically and semantically transparent but which still may be holistically stored and retrieved.

I have chosen to investigate phrasal patterns formed around one frequent noun: *hand*. One might argue that *hand* is a member of a group of "cardinal body nouns" similar to the "cardinal posture verbs" discussed by Newman & Rice (2004). Apart from being used for simple reference to a part of the body, *hand* occurs frequently with extended and metaphorical meanings. That body part nouns in general are a major source domain for metaphors is well-attested (cf. e.g., Lakoff & Johnson 1980; Goossens 1990; Gibbs et al. 2004) and similarly they are often the starting point of grammaticalization processes (Heine & Kuteva 2002). However, it is also of interest to see to what extent these words occur in formulaic sequences when they do *not* have figurative or nonliteral meaning.

2. Method and material

The method used is based on work by Stubbs (2002, 2007a and b), where he developed methods for corpus-based investigations of phrasal patterns in English, and has been described in some detail in Levin & Lindquist (2007) and Lindquist & Levin (2008 & forthcoming). The main points will be recapitulated here. Starting from a particular lexical item, in this case *hand* and its plural form *hands,* all recurring strings including this lexical item are retrieved from a corpus and are then submitted to analysis. The research is thus to a certain extent corpus-driven in the sense of Tognini-Bonelli (2001) in that there are no preconceived ideas about which strings or sequences will be encountered and in that the analytical categories as far as possible are based on the returns from the computer searches. The data were collected from the British National Corpus accessed by means of William Fletcher's (2003/2004) database *Phrases in English* (PIE), which includes all n-grams (identical strings of words) with a length between 1 and 8 words which occur 3 times or more in the corpus. Repeated searches were made, so that all n-grams from 2-grams to 8-grams and with the key word in different positions were detected. Figure 1 illustrates some 8-grams with *hand/hands* in different positions (where H stands for *hand/hands*, and + for any other word).

H + + + + + + +	
+ H + + + + + +	with **hands** clasped behind and palms facing inwards (6)
+ + H + + + + +	in the **hands** of the secretary of state (5)
+ + + H + + + +	
+ + + + H + + +	
+ + + + + H + +	take the book in your **hand** and repeat (3)
+ + + + + + H +	can be found on the left **hand** side (5)
+ + + + + + + H	

Figure 1. The extraction of 8-grams.

The lists of n-grams which were the result of these searches were then scanned manually for instances of phrases with internal semantic and syntactic integrity. Such recurring phrases are possible formulaic sequences, i.e., they are conceivably stored and retrieved holistically by some speakers (not necessarily by all).

Studies on individual lexical items require large corpora. For a relatively frequent word like *hand,* the BNC with its 100 million words proved sufficient. Using a standard corpus like the BNC makes it possible to control some aspects like national variety, genre and the spoken/written distinction, although in this particular case this was not a central concern. However, it is important to be aware of what has gone into the corpus one is using. The written component of the BNC (90 million words) contains 22 per cent "imaginative" writing, i.e., mainly literary texts. This explains why some phrases with *hand(s)*, e.g., *strong hands*, frequently occur in descriptions of erotic encounters typical of the popular romantic novels included in the corpus: *And the **strong hands** that slid round her waist were more real than any of her dreams*. It is therefore obvious and unavoidable that the kind of sequences found in the BNC will differ in distribution and probably also in kind from those that would be found in other corpora. Partington (1998: 107–108) suggests that one of the distinguishing features of genres is the types of metaphor that are found in them. One might paraphrase that and say that genres are distinguished by the type of formulaic sequence occurring in them.

In the following results section, a general overview of the search results for *hand(s)* will first be given, and then a closer study will focus on four sequences: *at hand*, *in hand*, *on hand* and *to hand*.

3. Results

3.1 *Hand(s)*

3.1.1 *Frequency of* hand(s)

In the 100-million-word British National Corpus, the lemma HAND occurs approximately 532 times/million words, thereby being the most frequent body term noun and the 26th most frequent noun overall (Leech et al. 2001). The singular form *hand* is twice as common as the plural form *hands*, with approximately 355 occurrences per million words against 177.

3.1.2 *Types of sequences with* hand(s)

The PIE program gives you an idea of the repetitive nature of naturally occurring language. In the 100 million words of the BNC there are for instance 514 different 6-grams with HAND that occur more than three times – and 6-grams are quite long strings of language. Of these many are chance occurrences of cut-out stretches of language without linguistic or semantic integrity like *hand and raised it to his*, *her hand to his lips and*, but many also have structure, as the following examples show. As expected, many of the n-grams describe routine, everyday actions that are carried out with the hands, at least as described in the fiction texts that are included

in the BNC. To some extent these are conventionalized ways of describing conventionalized actions, like the ones listed in Figure 2. Here strings with *his* and *her* have been combined.

ran a hand through his/her hair (10+12=22)
put a hand on his/her shoulder (11+6=17)
put his/her hand in his/her pocket (16+1=17)
laid a hand on his/her arm (11+3=14)
put a hand on his/her arm (6+6=12)
reached out and took his/her hand (1+8=9)
laid a hand on his/her shoulder (3+3=6)

Figure 2. Conventionalized actions described by formulaic 6-grams.

In passing, we can note that some of these actions are unisex, so to speak – both men and women run their hands through their hair – while some are typically male (putting one's hand in one's pocket) and others typically happen to males (*laid a hand on his arm*).[1] Such who-did-what-to-whom sequences could clearly be used in studies on gender stereotypes in language.

But we also find other types of phrases among the 6-grams, e.g., manner adverbials referring to physical acts like *with the back of his/her hand* (54 + 16 = 70) and *in the palm of his/her hand* (18 + 11 = 29). The latter phrase could of course also be classified as expressing position, i.e., something is placed in the palm of the hand. Another kind of position is expressed by *in the top left/right hand corner* (15 + 1 = 16) and *in the bottom left/right hand corner* (3 + 1 = 4). Normally, references to 'right' are slightly more common than references to 'left', probably as a consequence of more people being right-handed than left-handed. The surprising lack of balance between left and right here, however, can be explained by the fact texts on pages and computer screens begin in the top left hand corner.

Hand is also used for discourse organization as in *on the other hand it is/they are* (19 + 19 = 38) and *on the other hand there is/are* (16 + 21 = 37). Here the semantic change has gone from expression of location to discourse function. This is similar to what Traugott has described as "[m]eanings based in the external or internal described situation [changing into] meanings based in the textual and metalinguistic situation" (1989: 35). *Hand* is furthermore used metaphorically for 'help' in many 6-grams like *I'll give you a hand* (12).

Note that all the examples discussed in this section so far are chosen among the 6-grams. If one looks at shorter strings, the frequencies go up considerably.

1. Similarly, Lindquist & Levin (2008) found that *stamp(ing) his/her foot* was primarily used about females.

The 4-gram *on the other hand*, for instance, occurs 5,308 times in the corpus. This means that, astonishingly, 10% of all tokens of *hand* in the BNC occur in the phrase *on the other hand*. In fact, a quick check of 100 random tokens of *hand* from the BNC showed that 54% of the tokens occurred in formulaic sequences with nonliteral meaning like *on hand*, *to hand*, *on the one hand*, *give sb a free hand*, *cash in hand*, *hand in hand*[2], *try his hand* etc. This is support for Stubbs's suggestion that frequent words are frequent because they occur in frequent phrases (Stubbs forthcoming).

3.2 *At hand, in hand, on hand* and *to hand*

3.2.1 *The frequency of* at hand, in hand, on hand *and* to hand

It will not be possible to give the full "ecology" of such a common lemma as HAND in this short paper, so I will have to limit myself to one small area. I will look at four 2-grams which seem to overlap in quite intriguing ways, viz. *at hand*, *in hand*, *in hand* and *to hand*. These sequences show signs of lexicalization and entrenchment as fixed formulaic sequences in that they lack definite articles or possessive pronouns and in that they are in fact much more frequent than the regularly formed strings *at the hand*, *at his hand* etc. For instance, there were only 15 instances of *at the hand* (9 of which in turn were part of the 4-gram *at the hand of*, normally used figuratively meaning 'being (ill-)treated by'), compared to the 564 instances of *at hand*. All 19 instances of *at his hand* refer to looking or slapping etc. at a physical hand, or, in one instance, a glass of beer being placed close to a hand. Similarly, plural forms are much less frequent than singular. For instance, there were only 24 instances of *in hands*, 18 of which were part of the sequence *head in hands*, literally describing the posture of a person; another 4 expressed the meaning 'in somebody's care or possession'. Table 1 gives the frequencies for the 2-grams *at hand*, *in hand*, *on hand* and *to hand*.

Table 1. Frequency of *at hand, in hand, on hand* and *to hand*

at hand	564
in hand	1388
on hand	420
to hand	393
TOTAL	2765

2. This phrase occurs both with literal and non-literal meaning.

25. at hand. **a.** Within easy reach; near; close by. (Sometimes preceded by *close*, *hard*, *near*, *nigh*, *ready*.) **b.** Near in time or closely approaching. (Sometimes qualified as prec.)

29. in hand. **a.** *lit.* (Held or carried.) (---) **d.** In actual or personal possession, at one's disposal; (---) **f.** In process; being carried on or actually dealt with in any way. (---) **h.** ***in hand***: under control, subject to discipline. (Originally a term of horsemanship, cf. b.)

32. on hand, **upon hand**. **a.** In one's possession; in one's charge or keeping: said of things, or of work or business which one has to do. (---) **e.** At hand; in attendance (*U.S.*).

34. to hand. **a.** Within reach, accessible, at hand; †near, close by, close up, to close combat (*obs.*); into one's possession or presence. (See also *to come to hand*, 37a.) (---)

Figure 3. Selected meanings of *at hand, in hand, on hand* and *to hand* in the *OED on-line*.

The figure for *to hand* in Table 1 excludes infinitives and is extrapolated.[3] The total number of tokens of prepositional *at/in/on/to hand* including the extrapolated figure is 2765, which means that about 5% of all instances of *hand* in the BNC occur in one of these four phrases.

3.2.2 *The meanings of* at hand, in hand, on hand *and* to hand

The four sequences have a number of meanings ranging from the straightforwardly literal to the non-literal or figurative. The most relevant meanings given by the *OED on-line* for the four sequences are given in Figure 3.

As can be seen in the OED excerpts in Figure 3, there is a certain overlap in the meanings. In order to see how the sequences were used in the corpus, 200 random tokens of each 2-gram were first analysed to see what kind of objects etc. were described as being *at hand*, *in hand*, *on hand* and *to hand,* or with the formulation of the *OED*, what *at hand* etc. is "said of". The analysis will be presented in the next section.

---

3. The total figure for the string *to hand* in the corpus is 1,203. The figure 393 was extrapolated from a manual analysis of 600 examples, which yielded 196 prepositions. The option of using the BNC tags was not chosen, since a check of 50 tokens tagged as preposition + noun revealed that 14/50, or 28 %, were in fact wrongly tagged infinitives. Even if this small sample is not fully representative of the overall correctness of the tagging of this string, it indicates that the tagging can not be relied on. There are also tagging mistakes that go the other way: out of 50 random concordance lines tagged as infinitives, 3/50 or 6 % were wrongly tagged instances of prep + noun.

3.2.3 *Items described as being* at hand, in hand, on hand *or* to hand

The items described as being *at hand, in hand, on hand* or *to hand* were categorized as belonging to one of the categories Humans, Animals, Inanimate concrete items and Abstract items. An overview of the results is given in Table 2.

Table 2. Items described as being *at hand, in hand, on hand* or *to hand*

Item	at hand	in hand	on hand	to hand	Total
Humans	32	19	151	11	213
Animals	3	1	1	1	6
Inanimate concrete items	55	49	30	123	257
Abstract items	110	131	18	65	324
Total	200	200	200	200	800

Humans, Animals and Inanimate concrete items are relatively straightforward categories, while the category Abstract items subsumes a number of different subcategories which will be discussed below. As seen in Table 2, there is some specialization as regards the type of item that is referred to: *at hand* and *in hand* are similar in being used mainly about abstract things, *on hand* is used mainly about humans and *to hand* about inanimate concrete items. However, there is also considerable overlap, and, as will be shown below, several of the sequences can be used in identical contexts with seemingly identical meaning.

Humans

On hand stands out as the phrase which is most frequently used about humans, and the fine-grained corpus analysis gave a rather different picture from the one suggested by the *OED*. The 'in attendance' meaning, marked by *OED* as U.S., was by far the most common in this British corpus. In no less than 76% of the tokens of *on hand*, it is people who are *on hand*. Frequently the reference is to a 'specialist being available', as in (1).

(1) If pilots do get in trouble an instructor will be **on hand** to put them right. (CBF)

Other examples are about athletes, footballers and rugby players, who happen to be in the right spot to execute a good move, as in (2).

(2) Paul McGurnaghan's shot came back off the base of the post and David Eddis was **on hand** to hammer the ball into the net. (HJ3)

In a fair number of cases, the reference is to celebrities who are present at some occasion to perform some act, cf. (3).

(3) Believe it or not, Paul Newman is **on hand** to play the President and Susan Sarandon may play the first lady. (CK6)

With *at hand*, human reference is rarer but occurs with two main meanings, 'specialist available' as in (4) – (5) and 'in the vicinity' (with the word *close*) as in (6).

(4) […] the hard working Mr. Folten who is always **at hand** to offer advice and information on how to get the most out of your short visit. (EBN)

(5) […] she had been severely tempted to just throw in the towel and thumb through the Yellow Pages to find the nearest painter and decorator **at hand**. (JY5)

(6) Be prepared for this and ensure that you are close **at hand** with a reverse punch. (A0M)

Of the tokens of *in hand* classified as referring to humans, 16 are instances of the longer sequence *hand-in-hand*, which can be literal, as in (7) or figurative, as in (8).

(7) People strolled past without giving him a second look – couples **hand-in-hand**, families with pushchairs, groups of friends looking for the right spot for a picnic. (FS8)

(8) Good community care services work best where skilled professionals work comfortably **hand-in-hand** with unskilled staff, families, neighbours and voluntary organizations. (FYW)

Another 3 tokens of *in hand* referring to humans are really instances of *take in hand*, meaning 'take care/charge of', as in (9).

(9) For reasons too tedious to relate, the Pope is taken **in hand** by a natural healer […] (AKJ)

Among the 10 tokens of human reference with *to hand*, the 'specialist available' type as in (9) is the most common. However, some tokens refer to people who happen to be present, but are not necessarily experts, as in (10), and there was also one example of the sequence *bring to hand*, which means bring under control (11).

(10) It was surely an ideal situation for the police, with all the witnesses **to hand**, and even decent interview facilities. (C8D)

(11) "[…] introducing some of our ideas on personal training and discipline to ensure bringing the young men of the tribe **to hand** under our guidance in the early stages of their Moranhood." (C90)

Animals

Animals are occasionally referred to, as being experts – or perhaps rather a resource – as in (12), providing an attraction at close distance as in (13) or being put under control as in (14), which is an example of the original equestrian meaning of *take in hand*.

(12) The ferret is, of course, still on the line and remains close **at hand** on the surface near the whole. (BNY)

(13) Obviously I enjoyed having the opportunity to watch any nesting birds so close **at hand** […] (CHE)

(14) She took her mare **in hand** and clicked her tongue authoritatively. (HA2)

Being animate, but still treated by humans more or less as things, animals make up a small intermediate group between humans and the next group, inanimate concrete items.

Inanimate concrete items

In some cases with inanimate concrete items, as in (15) – (18), several of the phrases seem to be synonymous.

(15) Plasticine is useful to have **at hand** for propping up items of icing and marzipan while they are drying. (J11)

(16) Have an emergency tank **on hand**. (FBN)

(17) Have English mustard **to hand**. (CB8)

(18) […] but even the ordinary lay engineer, he looks to be able to do the job more efficiently, with the materials that he has **in hand** er and possibly introduce a new type of tool if he can get the proper material […] (GYV)

Example (18), from the spoken component of the corpus, is a bit unclear, and this was the only token where *in hand* was used in this sense. With concrete reference, the meaning of *in hand* was normally literal, 'with X held in the hand', as in (19) and (20), or a metaphorical extension of that meaning, as in (21) and (22). Note that in (19) and (22), the preposition *with* is present, whereas in (20) and (21), which represent the most common version of the construction by far, *with* is omitted.

(19) As the shadows lengthen, the men can be seen standing around with a pint of beer **in hand**, while mothers keep watchful eyes on the kids and catch up on the latest gossip. (A0V)

(20) *Paintbrush* **in hand**, Kylie recalls the beautiful things in life as she creates her own, very individual, landscapes […] (ADR)

(21) […] they were always prepared to swallow their pride and go, *cap* **in hand**, to the gentry for a few vital coppers. (G39)

(22) Or rather, they're going, but not with *cheque book* **in hand** and buying intentions in mind. (ACR)

Note also that in (19) there is an indefinite article on the held item – *a pint of beer* –, whereas in the more eroded versions of the construction the held items – *paintbrush, cap, cheque book* – do not take any form of determiner.

Abstract items

In the next section, the largest group, abstract items, will be discussed in greater detail. More than one third (39%) of the tokens were abstract, showing the result of a semantic development away from the original meaning of something concrete being situated at, in, on or close to a human hand. This group is more heterogeneous than the others; the labels given to the subgroups will become clearer in the discussion below. Table 3 gives a breakdown of the various abstract items referred to by *at hand*, *in hand*, *on hand* or *to hand*.

Table 3. Abstract items referred to by *at hand, in hand, on hand* or *to hand*

Item	at hand	in hand	on hand	to hand	Total
Task/issue/problem	36	28	9	5	78
Information	6	0	1	34	41
Help	28	1	1	2	32
Resources	11	1	3	15	30
Games up in sport etc.	0	22	0	0	22
Action in progress	1	22	0	0	23
Control	0	17	0	0	17
Ongoing activity	11	0	1	1	13
Improvements under way	3	0	2	0	5
Point in time	8	0	0	0	8
Tourist attraction	5	0	1	3	9
Money etc. in possession	0	5	0	2	7
Other	1	0	0	1	2
TOTAL	110	109	18	63	322

Task/issue/problem

One frequent type is reference to a task that someone has to carry out, an issue which is being discussed, or a problem that needs to be solved. These are common with *at hand* and *in hand* as in (23) – (24), and less common with *on hand* as in (25) and *to hand* as in (26).

(23) Making notes is the best way of keeping your mind on the task **at hand**. (EEB)

(24) However elaborate (indeed, contrived) this theorizing may be, it is not wholly adequate for the task **in hand**. (APH)

(25) Considering the possibly apocalyptic and doom-laden task we have **on hand** […] (CKC)

(26) They just got on with whatever task was **to hand**. (H7E)

The fact that all four sequences are used in the corpus with the same noun, *task*, without any clear difference in meaning, shows that there is variation,

probably between different speakers and possibly also in the same speaker. The meaning of the individual prepositions adds little to the meaning of the whole sequence. It is rather the sequence preposition + *hand* which is meaningful, and this makes it possible to exchange one preposition for the other in this particular context.

Information

I have chosen to include here also cases with reference to booklets, brochures etc., since the focus in these examples is normally on the content. Reference to information or information material is a specialty of the *to hand* construction. Example (27) is typical.

(27) It is therefore important to have all of this information **to hand** before beginning the installation procedure. (HAC)

With *at hand* and *on hand* there are only a few examples that can be put under this heading, cf. (28) and (29). With *in hand* there was none.

(28) Discipline means having an agenda, […] having your paperwork neat and easily **at hand** […] (EVF)

(29) […] for experience has shown how important it is to have these volumes **on hand** for cross-referencing with later works […] (B1P)

Help

Related to the meaning 'specialist available' is the abstraction 'help'. Just as with the concrete, human reference, this abstract meaning was most frequent with *at hand*, as in (30).

(30) But now help is **at hand**. You don't have to spend hours trudging around shops and in the kitchen to produce a perfect meal […] (HJ4)

The (contextual) synonyms *relief*, *remedies*, *assistance* and *tuition* also occurred, but *help* predominates, probably strengthened by the alliteration.

Resources

Under this heading has been collected other kinds of resources than specialists and help, as 'method' in (31) and 'favourable circumstances' in (32).

(31) In essence this was less a system than the resort by the Crown to whatever means came **to hand** for raising and controlling money. (EEY)

(32) Now although in evolutionary terms, given the amounts of genetic variability usually **at hand**, it is likely that such behaviour has been arbitrary in the required sense […] (CM2)

Games up in sports etc.

The second meaning unique to *in hand* is where the reference is to a sports team having played fewer games than the competitors, a tennis player having several serves left etc., as in the hopeful statement by a football supporter in (33):

(33) If we win them three games that we got **in hand** then we'll be up about sixth place. (KB4)

Action in progress

The meaning 'action in progress', as in (34) was also unique to *in hand*.

(34) I understand that an essential site survey has not yet been undertaken, and would ask that this be set **in hand** immediately. (HD2)

Control

The originally equestrian 'control' meaning given by OED for *in hand* occurred 17 times with abstract items, as in (35).

(35) If this is done frequently and perniciously it must be taken **in hand** and treated as a bad habit. (EEK)

Ongoing activity

This small category contains tokens which refer to something which is going on, usually in the vicinity, as in (36).

(36) Each time firing occurs close **at hand**, we all get down in the ditch […] (A61)

Improvements under way

In a few cases the sequence refers to improvements that have been or are believed/hoped to have been duly implemented and under way, as in (37) and (38).

(37) The release of five Western hostages in Lebanon during April generated intense, if premature, media speculation that a resolution of the whole hostage crisis was **at hand**. (HKT)

(38) Still, it is an election year and the nominally apolitical Fed is expected to rally round the incumbent president and dole out the credit with a liberal hand. But is there a recovery **on hand** to help? (AHJ)

This meaning is related to the 'control' meaning.

Point in time

At hand can be used to refer to a point in time, as in (39).

(39) The hour of Britain's total defeat was **at hand**, and Winston Churchill would soon abandon England to preside over the ruins from the safety of his Canadian dominion. (HWA)

Tourist attraction

Three of the sequences can be used to refer to attractions and activities that in some sense are available or attainable, as in (40) – (42).

(40) Another church visit is near **at hand**. St Laurence is on the corner of the Zeughausgasse immediately facing the cathedral precincts. (FTU)

(41) Volleyball, tennis and table tennis are **on hand** for the more energetic […] (AMW)

(42) During the summer months, the Wimbledon Tennis Tournament, the Oxford and Cambridge Boat Race and the Twickenham, Kingston and Richmond boat regattas are all **to hand**. (CJK)

Indeed, *at hand* as well as *on hand* and *to hand* are frequently listed under *available* in thesauruses (e.g., *Collins Compact English Thesaurus* 1993).

Money etc. in possession

Here we have the fixed sequence *cash in hand*, as in (43), but also the regularly constructed sequence *cash to hand* as in (44).

(43) Current assets are those that can be turned into cash at short notice, in addition to cash **in hand** or at the bank. (HRH)

(44) So whenever you have some spare cash **to hand**, pay it into Premier Savings and watch it grow. (EE0)

To conclude, the survey of abstract items shows that there is considerable overlap in use between two or three of the sequences in four of the categories: task/issue/problem, information, resources and tourist attraction. At the same time, there is clear specialization in three categories: games up, action in progress and point in time.

3.2.4 *Verbs used with* at hand, in hand, on hand *and* to hand

The four sequences occur either in verbless constructions or with a variety of verbs. There is a possibility that the choice of sequence was triggered by the verb used in the clause. The frequency for the most frequently occurring verbs are given in Table 4.

The figures in Table 4 show that 30% (240/800) of the tokens were verbless and that some verbs are used with all the sequences, although with varying frequencies. For instance, *on hand* is frequently used with forms of *be* in the construction *to be on hand*, while *at hand*, *to hand*, and especially *in hand*, are less commonly so. While some frequent verbs occur with all the sequences, others are specialized with only one or two, like *take* which only occurs in *take in hand* 'gain control over'. Similarly, a number of verbs like *go*, *walk*, *work*, *run* and *stride* exclusively or typically occur with the sequence *hand-in-hand* and verbs like *pass*

Table 4. Verbs most frequently used with *at hand, in hand, on hand* and *to hand*

Verb	At	In	On	To	Total
Verbless	81	95	20	44	240
Be	102	22	154	58	336
Have	8	18	17	34	77
Go		31		2	33
Come		1		29	30
Keep	5	1	3	10	18
Take		9			9
Pass				6	6
Lie				4	4
remain	1		2		3
Walk		3			3
Work		3			3

and *toss* occur with *from hand to hand*. To conclude this section, it seems that while in some cases any of a number of verbs will do, in other cases the verb itself is part of a longer sequence.

4. Grammaticalization or lexicalization?

Superficially, it might seem that the formulaic sequences studied in this paper have gone through a grammaticalization process from less grammatical to more grammatical, from the concrete, literal meaning in *at my hand*, to a more abstract, grammatical prepositional or adverbial meaning in *at hand*. However, I would argue that the process is rather a case of lexicalization of a kind that is close to what Brinton & Traugott (2005: 47) call "lexicalization as fusion" where a process of fusion results in decrease in compositionality (Brinton & Traugott 2005: 33) (see also Lehmann 2002). In particular, it is an instance of change from syntagm to lexeme. It can therefore be called a case of univerbation, where univerbation is taken to be the unification of a syntactic structure into a lexeme (rather than a word). A phrase like *at hand* still contains two words which make up a lexeme consisting of a multi-word unit. However this is not unlikely to be a stage in a development towards complete univerbation; the Swedish cognate, for instance, can be written either as two words or as a single word: *tillhands* (to+hand+GEN). As Brinton & Traugott point out (2005: 49), "[…] univerbations of older provenance often involve some degree of phonological reduction and are morphologically and semantically opaque, while those of more recent provenance may be relatively transparent both in form and meaning […]".

The sequences *at hand*, *in hand*, *on hand* and *to hand* agree with most of the traits given for lexicalization given by Brinton & Traugott (2005: 96–97) and do not contradict any of them. They have developed new meanings and syntactic forms over time (trait 1); the output is lexical and has to be learned by speakers (trait 4); they are fixed phrases (trait 5); they are the result of "gradual change in the sense that it is non-instantaneous, and proceeds by very small and typically overlapping, intermediate, and sometimes indeterminate steps" (trait 6); they begin to show signs of fusion (trait 7); and the components have lost their semantic compositionality (trait 8).

Whatever the end-point of this lexicalization process will be, it seems likely that old and new forms and meanings will coexist in the kind of layering described for grammaticalization by Hopper (1991). For instance, *in hand* may retain its concrete meaning in *with a glass of beer in hand* and the preposition *in* may be exclusively used in this sequence in this context, while the meaning of *in hand* in *the task in hand* may develop to become even more general with variable use of different prepositions. As far as solid spelling (indicating fusion) is concerned, a Google search (13 July 2007) threw up the following number of hits: *task athand* 241, *task inhand* 122, *task onhand* 51, *task tohand* 0. Some of these tokens are no doubt the result of sloppy typing on discussion forums and other technical accidents, but quite a few are from more reliable sources. The fact that there were zero hits for *task tohand* agrees with the results from searches for the fused forms *athand*, *inhand*, *onhand* and *tohand* on their own and in various other combinations, in that *tohand* is very much less frequent than the other three. It is not clear why this should be so, but it can perhaps be due to influence from the infinitive *to hand*. If there is ongoing univerbation, it is thus going at different speeds for the four sequences studied.

5. Conclusions

The partly corpus-driven method used in this study for retrieving n-grams including *hand(s)* from the British National Corpus gave as an output a large number of recurring strings, many of which can plausibly be considered to be formulaic sequences, at least for some speakers. The large number of recurring strings, even as long as 6-grams, illustrates that language is a mixture of repetition and creation, of drawing on stored sequences and constructing fresh strings by means of rules.

Concentrating on the four sequences *at hand*, *in hand*, *on hand* and *to hand*, it was demonstrated that these occurred with a number of different meanings, referring to humans, animals, concrete inanimate items and abstract items. The abstract meaning of the sequences are metonymic and metaphorical mappings of the literal body part meanings onto more abstract meanings.

In some uses the sequences overlapped so that two, three or even four different sequences could be used about the same item with no discernible difference in

meaning. This indicates that in these sequences, the prepositions *at*, *in*, *on* and *to* have lost most or all of their specific spatial meaning. The sequences are thus not only polysemous, they are also to some extent synonymous. In other uses, there is a strict specialization so that only one of the sequences can be used for a particular meaning.

The study of verbs used in connection with the sequences showed that 30% of the tokens were verbless. While some frequent verbs occur with all the sequences, others are specialized with only one or two. In some cases any of a number of verbs will do, whereas in other cases the verb itself is part of a longer sequence.

The sequences under study seem to be the result of a lexicalization process where they develop towards univerbation. This analysis is supported by the impression that they are often processed holistically without regard to the meaning of the preposition, and by the fact that they are beginning to occur written solidly on the World Wide Web.

In future studies it would be interesting to study sequences with other frequent body part nouns and other prepositions to see if similar developments can be found with these and if more wide-ranging generalizations can be made. In such studies it would also be desirable to take the historical development more into account than has been possible in the present paper.

References

Brinton, Laurel J. & Elizabeth Closs Traugott. 2005. *Lexicalization and language change*. Cambridge: CUP.

Collins compact English thesaurus in A–Z form. 1993. Glasgow: HarperCollins

Firth, J.R. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis* [Special Volume, Philological Society], 1–32. Oxford: Blackwell.

Fletcher, William. 2003/4. *PIE: Phrases in English*. http://pie.usna.edu.

Gibbs, W. Raymond Jr., Paula Lenz Costa Lima & Edson Francozo. 2004. Metaphor is grounded in embodied experience. *Journal of Pragmatics* 36: 1189–1210.

Goossens, Louis. 1990. Metaphtonomy: The interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics* 1: 323–340.

Heine, Bernd & Tania Kuteva. 2002. *World lexicon of grammaticalization.* Cambridge: CUP.

Hopper, Paul. 1991. On some principles of grammaticalization. In *Approaches to grammaticalization,* Vol. 1, E. Closs Traugott & B. Heine (Eds), 17–35. Amsterdam: John Benjamins.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar*. Amsterdam: John Benjamins.

Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago IL: The University of Chicago Press.

Leech, Geoffrey, Paul Rayson & Andrew Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus.* London: Longman.

Lehmann, Christian. 2002. New reflections on grammaticalization and lexicalization. In *New reflections on grammaticalization*, I. Wischer & G. Diewald (Eds), Amsterdam: John Benjamins.

Levin, Magnus & Hans Lindquist. 2007. Sticking one's nose in the data. Evaluation in phraseological sequences with *nose. ICAME Journal* 31: 63–86.

Lindquist, Hans. 2007. Review of: Römer, Ute: *Progressives, patterns, pedagogy. A corpus-driven approach to English progressive forms, functions and didactics. International Journal of Corpus Linguistics* 7(2): 119–131

Lindquist, Hans & Magnus Levin. 2008. Foot and Mouth: The phrasal patterns of frequent nouns. In *Phraseology: An interdisciplinary perspective*, S. Granger & F. Meunier (Eds), 143–158. Amsterdam: John Benjamins.

Lindquist, Hans & Magnus Levin. Forthcoming. The grammatical properties of recurrent phrases with body part nouns: The N_1 *to* N_1 pattern. To appear in *Exploring the lexis-grammar interface*, U. Römer & R. Schultze (Eds), Amsterdam: John Benjamins.

Mair, Christian. 2006. *Twentieth-century English. History, variation and standardization.* Cambridge: CUP.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon.

Newman, John & Sally Rice. 2004. Patterns of usage for English sit, stand, and lie: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics* 15(3): 351–396.

Oxford English dictionary on-line. www.oed.com.

Palmer, Harold E. 1933. *Second interim report on English collocations.* Tokyo: Kaitakusha.

Partington, Alan. 1998. *Patterns and meanings. Using corpora for English language research and teaching* [Studies in Corpus Linguistics 2]. Amsterdam: John Benjamins.

Schmitt, Norbert, Sarah Grandage & Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In *Formulaic sequences. Acquisition, processing and use* [Language Learning & Language Teaching 9], N. Schmitt (Ed.), 127–151. Amsterdam: John Benjamins.

Sinclair, John. 1991. *Corpus concordance collocation*. Oxford: OUP.

Sinclair, John. 1999 A way with common words. In *Out of corpora*, H. Hasselgård & S. Oksefjell (Eds), 157–79 Amsterdam: Rodopi.

Sinclair, John. 2003. *Reading concordances. An introduction.* London: Longman.

Stubbs, Michael. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, Michael. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7(2): 215–44.

Stubbs, Michael. 2007a. An example of frequent English phraseology: Distributions, structures and functions. In *Corpus linguistics 25 years on*, R. Facchinetti (Ed.), 89–105. Amsterdam: Rodopi.

Stubbs, Michael. 2007b. Quantitative data on multi-word sequences in English: The case of the word *world*. In *Text, discourse and corpora,* M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert, 163–189. London: Continuum.

Stubbs, Michael. Forthcoming. Quantitative data on multi-word sequences in English: The case of prepositional phrases. Lecture given at the Berlin-Brandenburgische Akademie der Wissenschaften, 3 November 2006.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.

Traugott, Elizabeth Closs. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 65: 31–55.

Wasow, Thomas. 2002. *Postverbal behaviour*. Stanford CA: CSLI.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.

The embodiment/culture continuum

A historical study of conceptual metaphor*

James J. Mischler, III
Oklahoma State University

Abstract

Cognitive Linguistics accepts as a fundamental principle that embodied experience and culture both influence the cognitive conceptualization of meaning in language; however, most studies focus on the influence of embodiment. Diachronic studies are useful to show the effect of cultural models on conceptualization. The current study collected samples of metaphors of the spleen (e.g., "He vented his spleen") from 19th century English popular

*I would like to thank Carol Moder and Nick Ellis for their insights on the study topic, which sharpened my thinking during the writing of this paper. I also appreciate the work of the anonymous reviewer, who provided excellent notes, corrections, and advice on the paper draft. Any errors are my own.

Several academic organizations contributed funding and other resources to the study, including the Robert Glenn Rapp Foundation, Oklahoma City, Oklahoma USA, and the Graduate College and the Department of English at Oklahoma State University. The financial support of these organizations is greatly appreciated.

magazines to investigate the relative contributions of cognition and culture on metaphor instantiation. The results showed that culture was isomorphic with embodied experience in the data. Based on the results, an embodiment/culture continuum is proposed, within which different conceptualizations vary in their content on the two dimensions. Usage-based models of language provide an explanation for the study results.

1. Introduction

Cognitive Linguistics (CL) is founded on the principle that the human ability for language is the product of general cognitive processes of the brain, especially the **cognitive conceptualization** of human experience in the physical body (Langacker 1987). Repeated physiological and sensorimotor movements of the limbs, face, and internal organs over time provide the cognitive structuring, or conceptualizations, that form the experiential basis for thought and language. Thus, repeated **embodied experience** leads to cognitive structures that are used to interpret later experiences. Certain types of experiences (e.g., breathing) are so basic that they are universal, creating conceptualizations that are instantiated across languages and cultures (Lakoff & Johnson 1999). Languages reflect the presence of theses universal experiences; for example, many languages conceptualize anger as a physical feeling of pressure in the body (Kövecses 2005).

The model of the relationship between embodiment and language in CL is not a closed system, however. Other factors, including cultural knowledge and social interaction, are also viewed to play an important role. Lakoff, in his work on **conceptual metaphor** (1987), posited Idealized Cognitive Models (ICMs) as a type of gestalt which is organized by several different types of cognitive "structuring principles" (Lakoff 1987: 68). Crucially, ICMs vary from culture to culture, as Lakoff shows in the differences between the English ICM for a calendar week (i.e., seven days based on the movement of the sun) and the same ICM in Balinese (i.e., ten separate "cycles" of day names). The names of the days in each language are a result of the operation of the ICM (68–69). In sum, **cultural models** of human experience influence conceptualization and language structure.

Though cognitive linguists acknowledge cultural models as important factors in conceptualization and linguistic meaning, research has tended to investigate the physiological and sensorimotor bases of conceptualization, often as a result of theoretical work which has the same focus. In response, William Croft, in a forthcoming article (available at http://www.unm.edu/~wcroft/WACpubs.html) goes so far as to argue that the fundamental principles of CL are "too much 'inside the head'. In order to be successful, cognitive linguistics must go 'outside the head' and incorporate a social-interactional perspective on the nature of language" (Croft

forthcoming 1; see also Gibbs 1999; Kövecses 2005). Thus, social and cultural knowledge is isomorphic with physical experience of the body in conceptualization. CL can account for variations in conceptualization which do not follow universal physical experience. Since cognition and culture are linked in CL, changes in conceptualization may be the result of changes in cultural models. The purpose of the current study is to apply Croft's idea to metaphor in English.

Metaphor has been recognized as a type of formulaic language. In a recent book-length analysis of many types of formulaic sequences (Wray 2002), metaphor is included as a type. Wray reviews definitions of **idiom** which include the features of non-compositionality (i.e., a "frozen" form) and wholistic meaning (i.e., meaning derives from the formulaic sequence as a whole). However, she argues that some idioms are flexible in terms of compositionality: "fluidity allows for them to be componential on one occasion and entirely wholistic on another" (Wray 2002: 57). Compositionality can become a permanent structural feature of a formulaic sequence, resulting in "flexible slots … which can be filled with semantically-appropriate words or phrases" (Schmitt & Carter 2004: 5). Metaphoric expressions also may have slots; thus, like other types of formulaic language, metaphors vary on a continuum of compositionality, ranging from wholistic to highly compositional (Clausner & Croft 1997). In sum, in this paper metaphoric expressions are considered to be a type of formulaic sequence.

2. Historical studies of conceptual metaphor

In order to study cultural models and their effect on conceptualization, a diachronic design was employed for the current study. The study of historical forms of conceptual metaphor can show more clearly the interplay of language and culture, for several reasons. First, cultural models may structure a conceptual metaphor diachronically and aid in the spread of the conceptualization across languages. Geeraerts & Grondelaers (1995) found a wide range of evidence in language and art that anger metaphors in English and Dutch may be the result of the historical Four Humors cultural model of medicine, which was a prominent cultural practice in Europe and Great Britain during the Renaissance and Enlightenment periods. Geeraerts and Grondelaers acknowledged that the historical linguistic forms may have been the result of universal experience in the body, as conceptual metaphor theory holds, but their evidence does allow for the possibility that the cultural model may have been the basis for the conceptual metaphor. Similarly, Gevaert (2002), in a study of conceptualization in Old and Middle English, found that present-day anger metaphor (e.g., "his blood boiled") became prominent in English when the Four Humors cultural model spread the concept of heated bodily fluid from France to Great Britain

in the mid-15th century. MacArthur (2005) found that conceptual metaphors of horse riding in Spanish, arising historically in the aristocratic cultural sub-group, were spread to other sub-groups in Spain as a result of the localized, culturally-based authority and influence of the aristocratic social class, rather than through universal embodied experience. These studies indicate that cultural models may influence in some way the development of conceptual metaphors over time.

Second, over time, cultural models may change the meaning or cultural significance of a conceptual metaphor. Gevaert (2002), mentioned previously, found that the cognitive conceptualizations of anger changed from AD 850 to 1450. Early in that period, the concept of "heat" comprised 1.58% of the data; by the 15th century, heat comprised 3.64%, more than double the earlier rate. As was stated earlier, Gevaert attributes the change in the 15th century to the rising popularity of the Four Humors model. Koivisto-Alanko & Tissari (2006: 210), in a study of metaphors reason and emotion, found that particular mappings changed over time in their semantic meaning or cultural content. For example, the meaning for "wit" changed from "mental manipulation" to personification of a learned person to "valuable commodity." Fear as an emotion changed in its cultural value from "negative value" to "valuable commodity" several times. Thus, culture was found to have an important effect on the structure and meaning of the conceptual metaphors.

The value of historical study to understand the relation between language and culture has been noted by many linguistic researchers, including those in cognitive linguistics. Sweetser (1990) points out that synchronic forms are the result of diachronic processes. Bybee (1988), discussing Greenberg's (1957) research program for general linguistics, concludes that "synchronic states must be understood in terms of the set of factors that create them. That is, we must look to the diachronic dimension …" (1988: 351). Allan (2006) states that the need for historical data on metaphor is an imperative for researchers: "Many of the metaphors pervasive in everyday language are products of their time, and cannot therefore be accounted for without reference to culture" (Allan 2006: 175). It is non-controversial in linguistics that language forms have historical meanings and uses that affect the development and use of the synchronic form. The current study accepts this basic principle as the basis for investigating historical metaphor.

To summarize this section, diachronic studies show that conceptual metaphors reflect particular cultural models, are spread via those models and social interaction, and change over time as a result of cultural influence. If universal cognition is the basis of metaphor, then change should be temporary and/or minor in its impact on semantic meaning. The long-term nature and broad scope of cultural influence on conceptual metaphor in historical research calls into question current theory.

The goal of the present study is to explore further the role of cognition and culture in the instantiation of one historical metaphoric expression in natural

language data. Enfield (2002) has pointed out the need for methodological rigor in studying the effect of culture on language. He recommends the inclusion of data, both linguistic and non-linguistic, outside of the language phenomena under study, in order to provide corroborating evidence for the effect of culture on language. This approach will be employed in the current study. By adopting appropriate data collection and procedures for analyzing the data, the results of the current study can deepen our understanding of the interplay between cognition, language, and culture.

3. Spleen metaphors of anger

To investigate the influence of cognition and culture on metaphor, I chose another body-based conceptual source domain of human emotion – the spleen. There are several reasons for this choice. First, the spleen is mapped to anger in English. The *Oxford English Dictionary Online* (hereafter, OEDO; available at http://dictionary.oed.com.) states that *spleen* as a noun can signify a "[a] sudden impulse; a whim or caprice"; "[h]ot or proud temper"; "[v]iolent ill-nature or ill-humour; irritable or peevish temper"; or, "[a] grudge; a spite or ill-will" (OEDO, n.d.). Coupled with the transitive verb *vent*, which can mean "To let loose, pour out, wreak (one's anger, spleen, etc.) on or upon a person or thing" (OEDO), a typical form is *He vented his spleen*, a metaphoric expression which refers to verbalized emotion, particularly anger, irritation, or sarcasm. All of these meanings are now archaic in terms of present-day use; however, they show that historically the spleen has served as a source domain for metaphoric expressions of emotion in English.

Another reason for choosing the spleen is that, historically in Western culture, the organ was associated specifically with the Four Humors cultural model. As mentioned previously, a study of the influence of the Four Humors in anger metaphor in English and Dutch (Geeraerts & Grondelaers 1995) indicated that the cultural model influenced significantly the conceptual metaphor employed in historical texts. In addition, Gevaert's (2002) study of Middle English indicated that the conceptualization of heat increased when the Four Humors model was introduced in Great Britain in the 15th century. The spleen had a prominent role as one of the four major organs described in the Four Humors; therefore, the possibility exists that the cultural model influenced the metaphor.

4. The Four Humors model

To understand the possible role of culture on spleen metaphors, background data on the Four Humors model was collected. The Four Humors was employed as a sci-

entific, empirically-based model of human health for over 2,000 years in Western culture, up until the mid-nineteenth century (Nutton 1993). The system informed the diagnosis and treatment of all types of illnesses, both physical and mental.

The four humors (or, "fluids") included blood, phlegm, yellow bile, and black bile. The bodily fluids ideally needed to be balanced for a person to be in good health; however, perfect balance was viewed as a rare phenomenon, and one of the humors usually dominated a particular individual. The dominant humor was believed to be produced in quantities in excess of the body's needs (in the model, all four humors provided health benefits), with the result that the person became prone to particular physical illnesses and to a particular psychological temperament. The goal of medical treatment in the Four Humors model was to restore the balance of the fluids, and thereby improve physical and psychological health (Nutton 1993).

The humors were also linked to other natural phenomena, including the four qualities. Each of the fluids possessed two of the qualities: "Blood, like air, had the qualities of heat and moisture; phlegm, like water, those of coldness and moisture; yellow bile, like fire, those of heat and dryness; and the atrabilious humor [black bile] those of cold and dryness, like earth" (Hoeniger 1992: 102–103). These qualities were used to determine the type of treatment to give for a particular disease and its symptoms. For example, a disease caused by cold, dry black bile would receive treatments which possessed the opposite qualities, namely heat and wetness (Ackerknecht 1982). The four qualities were important elements in the diagnosis and treatment of diseases.

The four humors were in turn linked to four human physiological organs – the heart with blood, the brain (or stomach) with phlegm, the liver with yellow bile, and the spleen with black bile. Both physical and psychological conditions could be treated through the system. The spleen was believed to produce *melancholy*, which was viewed as both a personality type and a medical condition. Excess black bile also was linked to certain physical traits, a specific personality type, and medical conditions. For example, a tall, lean body, a desire for solitude, social behavior characterized by shyness, and a tendency toward sadness or anger, were all viewed as symptoms of excess black bile. The link between a humor and a bodily organ in the Four Humors model was extrapolated to explain a wide range of physical and emotional phenomena.

Though black bile was cold, a form of "hot" bile, namely *melancholy adust*, the result of burning black bile until the fluid was destroyed, was thought to cause extreme forms of anger, including madness. In sum, the spleen and black bile were important features of the Four Humors model. The features were categorized under the "melancholic" personality type, which included particular physical traits and a specific set of physical and mental illnesses.

5. Method

5.1 Materials

The metaphor data were collected from two Internet digital corpora of historical British and American magazines. One site is the *Internet Library of Early Journals* (hereafter, "ILEJ"), located at http://www.bodley.ox.ac.uk/ilej/, a cooperative project initiated by the universities of Birmingham, Leeds, Manchester, and Oxford to digitally preserve early British magazines from the mid-eighteenth century to the mid-nineteenth century. The second site is *The Nineteenth Century in Print* (hereafter, "NCP"), located at http://lcweb2.loc.gov/ammem/ndlpcoop/moahtml/snchome.html, a collaborative effort between Cornell University, the University of Michigan, and the U.S. Library of Congress, to preserve American popular magazines. Both collections are searchable by keyword using a Web interface, and the results returned include one or more magazine pages (from the original source) for a single instance of the keyword, thus providing the contextual information needed for close analysis. This last feature is important because studies of compiled corpora have been criticized for the lack of contextual information provided by corpus search programs (Hunston 2002). The corpora selected for the current study addressed this limitation by providing the full context in which each keyword instance was situated.

The specific texts chosen for the data collection were two English language magazines published from 1844 to 1863 – *Blackwood's Edinburgh Magazine* from ILEJ and *Littel's Living Age* from NCP (hereafter, *BEM* and *LLA*, respectively). The magazines and the specific time frame were selected because the digitized volumes for both publications are complete for the continuous 20-year period.

5.2 Data collection

The full text of the digitized volumes was searched using the keyword *spleen*, *and* the first 100 instances of the keyword for each magazine were accepted for analysis. 171 cases resulted from the keyword search procedure. 42 duplicate instances, which were likely artifacts of the corpora search algorithms, were eliminated; the remaining 129 cases comprised the study sample. Non-metaphorical uses of the keyword were also excluded, including medical references (20 cases), leaving 109 spleen metaphor cases.

The metaphoric expressions employed the word *spleen* as the source domain, with a target domain (either present lexically or implied contextually) that signifies an expression of emotion. In addition, the context in which each instance of the keyword appeared was carefully read and evaluated to determine the target

domain; only samples categorized within the target domain of human emotion were accepted for analysis. The sample below shows the problems of interpretation inherent in some of the collected samples.

> *Fair Saint George*
> *Inspire us with 'the spleen of fiery dragons'!*
> *Upon them! Victory sits on our helms.*
> (*BEM,* January, 1858; p. 131)

The metaphor targets courage, which is not an emotion; emotion types are implied rather than explicitly identified in the text. In addition, the amount of context available is inadequate to categorize the target domain of the metaphor clearly. In all, 12 cases were eliminated due to the unclear meaning of the metaphor within a given context. A final total of 97 cases of spleen metaphors of emotion were collected from the original 129; metaphoric expressions of emotion accounted for 75.2% of the collected keyword instances. A summary of the data collection procedure is shown in Table 1.

Table 1. Keyword instances, excluded cases, and study cases for the selected magazines

Magazine	Instances*	Excluded	Study Cases
BEM	68	18	50
LLA	61	14	47
Totals	129	32	97

*Excluding duplicate cases

5.3 Data analysis

In the data analysis phase, the 97 collected spleen metaphor cases and their original context were compared to the six properties of embodiment from Lakoff & Kövecses's (1987) folk model of human physiology. The model includes the following six cognitive conceptualizations of embodied experience: the CONTAINER image schema; container pressure; fluid; heat; the heat scale; and visible physiological effects (i.e., skin redness, bodily agitation, and impairment of vision). These six conceptualizations are mapped onto the target domain of ANGER, forming the primary conceptual metaphor ANGER IS HEAT.

The six properties conceptualize different aspects of the human experience of the physical body during an instance of anger. The CONTAINER is a schematic representation of the human body as a container of bodily fluids, and heat is the experience of the rise in body temperature when angry. Increased heat leads to the experience of pressure; that is, there is a tendency to suppress the expression

of emotion. The heat scale is a cognitive representation of the relative rise and fall in heat and pressure during anger – heat increases as anger increases, and heat decreases as anger decreases. Finally, visible physiological effects are the result of a high increase in anger: as anger increases to an extreme level, the body shows certain physical manifestations, such as the skin around the face and neck turning red, arms and legs shake, and visual acuity is impaired. Lakoff & Kövecses (1987) shows that the properties are employed systematically in English metaphoric expressions, such as *His blood boiled,* and the properties are the result of human experience in the physical body. Thus, language is inextricably entwined with the cognitive conceptualization of everyday experience.

Lakoff & Kövecses' analysis argued for a universal, cross-linguistic link between human bodily experience and the cognitive conceptualizations that create metaphoric expressions in language. Each sample was analyzed for all six properties. In the Cognitive Linguistics' view, the properties do not all have to be present in each sample because no member of a cognitive category has all of the features of the category (Lakoff & Johnson 1999), but evidence of one or more is necessary to provide evidence for universal embodiment. It also must be pointed out that the purpose of the comparison procedure was to facilitate the identification of embodied experience in the data; the method does not imply that a conceptual metaphor motivating the spleen samples is related to the ones described by the Lakoff and Kövecses folk model. The procedure was designed to facilitate the systematic identification of specific aspects of embodied experience in the data.

6. Results

6.1 Data examples

The following section shows examples of spleen metaphors and their common characteristics. Example 1 displays the container, pressure, and fluid properties described by Lakoff & Kövecses (1987).

(1) *In short, altogether he is put out, and he vents his spleen on the swans, which follow him along the wave as he walks along the margin, intimating either their affection for himself, or their anticipation of the bread crumbs associated with his image …* (*BEM,* January, 1859, 1)

The context of the sample indicates the emotion of anger in the words *put out*. The spleen metaphor is in its prototypical form, *He vents his spleen*, and is employed to refer to the verbal expression of anger.

The properties of pressure and fluid are instantiated in the word *vents*. The OEDO states that the verb can mean "to discharge, eject, cast or pour out (liquid, smoke, etc.); to carry off or away; to drain in this way … Said usually of the containing thing, but sometimes of 'the force' or means by which outlet is given." Force indicates pressure on the fluid. An alternative meaning of the verb applies to " … persons, animals, or their organs: To cast out, expel, or discharge, esp. by natural evacuation; to evacuate" (all above quotations from OEDO). Therefore, in (1) the fluid is vented from the spleen to reduce pressure, and this conceptualization denotes a verbal expression of anger.

The following example also shows pressure in the container.

(2) *There is one fallacy, however, still current against woman, which we must take this public opportunity of renouncing. A certain ungallant old father, soured by the circumstances of his lot, relieved some of his spleen by defining women [Greek translation] – an animal that delights in finery …* (*LLA,* May 1847, 337)

The use of the verb form *relieved* indicates that the spleen is under pressure, and also that the pressure can be decreased by direct action; OEDO describes the meaning as "[t]o give (a person, part of the body, etc.) ease or relief from physical pain or discomfort." *T*he social situation indicates that the man is "taking out his anger" concerning his own disappointments in life and directing his emotion against other people, in this case, women in general. Anger is a common emotion displayed in spleen metaphors.

The next example shows the negative consequence of failing to relieve the pressure on the spleen.

(3) *And this interesting piece of geographical, and geological, and hydrographical meditation makes part in a "burst of indignant spleen" which is to go near to "annihilating" Man from the face of the Globe!* (*BEM*, August 1854, 201)

Burst denotes a sudden, explosive destruction of the container as a result of extreme pressure; the violent force of the explosion is displayed in the metaphorical destruction of the Earth. This property of destructive force is similar to the one displayed in the anger metaphors analyzed by Lakoff & Kövecses (1987).

The idea of bursting was also expressed in the data through various synonyms, such as *ebullition*, in (4).

(4) Swift *calls Ruvigny* '*a deceitful, hypocritical, factious knave, – a damnable hypocrite of no religion;' but this is a mere ebullition of spleen, such as was common with Swift against a Whig opponent.* (*LLA,* November 4, 1854, 495)

Ebullition in the OEDO is defined as "[t]he process of boiling, or keeping a liquid at the boiling point by the application of heat." The presence of a word that indicates heat is significant, since that is congruent with Lakoff and Kövecses's

findings. However, the result also is consistent with the Four Humors model; recall that the spleen produced a range of emotions in the medical model, including hot anger (i.e., *melancholy adust*). Whether the property originates in embodiment or culture is difficult to discern; a longitudinal study would be needed to research the question, which is beyond the scope of the current investigation. In the 97 cases collected, four were found to use words indicating heated liquid or heated containers; *ebullition* was used in three cases, or 3.1% of the total cases. The feature of heat is a rare one in spleen metaphor, yet this result is consistent with the cultural model.

All of the spleen samples include the fluid property, but the qualities of the fluid are often different from the fluid found in Lakoff and Kövecses's data – though always under pressure, spleen fluid is generally unheated, with a few exceptions, as noted in (4). Example (5) provides further evidence of the absence of heat and steam.

(5) *… he passed the next ten years of his life agreeably enough, if not contentedly. He found a vent for his spleen in the practice of political journalism, and it was during this period that many of his finest works were written …*
(*LLA,* September 12,1863, 518)

Heat and steam result in rising fluid and physiological effects, such as skin redness and bodily shaking; those characteristics are not found in (5) or in any of the spleen samples, indicating that heat and steam are not present. Unheated fluid is a typical characteristic of the spleen data.

Consequently, the heat scale found in Lakoff and Kövecses's data was not present in the spleen data, either; that is, an increase in emotional intensity is not the result or the cause of an increase in temperature, as in (6), below.

(6) *Those who knew him best say that, about this time, his temper became horribly soured. He never had been very agreeable in the servants' hall, but now he was snappish and morose … But as he durst not quarrel with Gray, he resolved to vent his spleen upon somebody else, and to his own infinite misfortune, selected Protocol as the victim.* (*BEM,* February 1853, 169)

The emotional intensity increases as the passage progresses – the man is first disagreeable, then sour, then morose, and finally decides to vent his increasing anger by quarreling with another man, named Protocol. The fluid pressure increases to the point that venting is desired, yet the fluid temperature does not change; therefore, the increase in temperature that signals increased anger in the Lakoff and Kövecses samples is not present. It is notable that a heat scale is not present in the cases in which heated fluid is present, either (see sample 4). Fluid in the spleen metaphors is usually unmarked for heat, and in all cases, greater intensity of

emotion is not the result of increases in heat and steam. Instead, the bursting of the container is caused by excessive fluid volume in the spleen, exerting pressure on the container.

The sixth property of Lakoff and Kövecses' model, visible physiological effects, is also absent in the spleen metaphors. In fact, none of the collected samples display the visible effects identified by Lakoff and Kövecses. Two cases illustrate the lack of physiological effects. First, in (7), black bile affects the mind, rather than the body.

(7) *That those who would die for and with each other in the hour of peril, are but too apt to misuse the hour of prosperity in conceiving groundless jealousies, in attributing undue importance to passing bursts of spleen and petulance, in mutual and self-torment. It is the original sin of man to take advantage of the absence of important evils to magnify in his imagination those of minor consequence …*
(*LLA,* October 19, 1844, 670)

The result of the metaphoric *bursts of spleen* is mental suffering for both the person who expresses spleen anger and those around him or her. Mental health is implicated in (7) through the use of emotion words (e.g., *petulance*) and words linked to internal thoughts (e.g., *mutual and self-torment* and *magnify in his imagination*); however, visible physiological changes are not displayed. Thus, the bursting of the spleen (i.e., a violent expression of emotion) occurs suddenly and without warning, due to the lack of visible signals of anger.

As further evidence of the psychological effects of spleen anger, example (8) shows the extreme effects on the sufferer's mental state.

(8) *When labouring under a bad attack of the spleen – so said our volatile and veracious neighbours – the Englishman felt his life to be a burden to him. Nothing but family considerations … prevented him from blowing out his brains with a pistol, or effectually ridding himself of his woes by plunging into the muddy torrent of the Thames.* (*BEM,* September 1861, 302)

The text indicates that prolonged or intense exposure to black bile results in extreme negative thoughts, including the consideration of suicide. Similar to (7), visible physiological *sensations*, such as skin redness, are not manifested in spleen metaphors. The effects are psychological, not physiological.

6.2 Summary of the results

The collected samples of spleen metaphors are systematic and consistent in their instantiation: the CONTAINER is the spleen, and the container is under pressure; there is fluid in the spleen, and the fluid is unheated; the expression of anger typically occurs suddenly and without warning, due to the lack of visible physical effects, and the

resulting behavior can have severe emotional and psychological consequences, such as depression and suicidal thoughts, both for the person expressing the emotion and for others present at the time of the emotional outburst. The results show that spleen metaphors are markedly different from the anger metaphors analyzed by Lakoff and Kövecses. In addition, it is unclear what, if any, embodied experience motivates the linguistic expressions, considering the lack of physiological content in the samples. As stated earlier in this paper, though evidence of the six properties are not required in each sample to provide evidence of embodiment, the consistent absence of their conceptualization in spleen metaphors is an indication that some other factor, such as the Four Humors model, is providing a significant portion of the conceptual content.

In addition, the Four Humors model correlates with the characteristics of spleen metaphor. Specific aspects of the Four Humors view of the spleen that correlate with the content of the spleen metaphors include the absence of heat and the heat scale, the lack of visible physical symptoms such as skin redness and bodily agitation, and the focus on psychological mental states rather than physical sensations.

7. Discussion

Assuming that spleen metaphors were conventional forms during the historical period under study, what can account for their existence and systematic instantiation? The answer suggested by the study results is that culture affects the conceptualization of experience. As a result, shared cultural views of the human body informed the systematic use of the metaphor in the nineteenth century for speakers in the speech community who knew the cultural model. In this way, cultural knowledge is isomorphic with embodied knowledge, as Croft's (forthcoming) proposal implies.

The "isomorphic" view is somewhat simplistic, however. Based on the results of the current study, embodiment and culture can be posited as two ends of a continuum; across different conceptualizations, the metaphor content varies between the two factors, depending on the origin, historical development, current cultural beliefs and values, pragmatic meaning in situational context, and the experience of an individual speaker. Historical and cultural factors determine the content that is available for use (including content that is no longer available due to historical changes in form and meaning and changes in the cultural beliefs of the speech community), and synchronic factors, including the speaker's creativity and individual experience, determine the current use of the conceptualization in social interaction. Thus, spleen metaphors are closer to the "culture" end of the continuum, having been formed initially by the influence of the Four Humors model hundreds of years ago. Lakoff and Kövecses' anger metaphor is

closer to the "embodiment" end, in large part formed by everyday experience in the human physiological body. Other conceptualizations can be placed at various points along the continuum, depending on their specific content, historical origin, and current use.

In sum, the embodiment/culture continuum establishes the semantic **range** for the conceptualization (see Croft forthcoming: 24), while also providing open-ended choices for synchronic use in context. The specificity and flexibility of the continuum can explain the relative stability of conceptualization over time, the systematic use of the concept, and the synchronic creativity of linguistic forms.

The theoretical view of conceptualization as a continuum is consistent within the **usage-based model of language**. Usage-based models (Barlow & Kemmer 2000; Bybee 2001; Tomasello 2003) view language as a result of multiple input factors, including cognitive, cultural, and interactional processes, such as conceptualization, knowledge of the speech community and the world at large, and the social factors which govern language use, including the situational context and frequency of use. This multiple-factor view is flexible enough to explain language forms which are based in culturally-sanctioned concepts, such as spleen metaphors licensed by the Four Humors model, as well as forms derived from general cognitive principles, such as Lakoff and Kövecses's anger metaphors. Usage-based models of language can account for both embodiment and culture and their mutual interaction.

8. Conclusions

The current investigation has shown that, for spleen metaphors, the Four Humors cultural model correlated with the specific characteristics of the metaphor. The result suggests that the effect of culture on the linguistic expression is isomorphic with embodied experience as a factor in conceptualization. Based on the results of the study, an embodiment/culture continuum is proposed. The concept can explain both the stable structural features and the synchronic, creative use of conceptualizations. In addition, the continuum fits the multiple-factor view of cognition and language found in the usage-based model of language.

Further research is needed on spleen metaphors to investigate the form in more depth – longitudinal study is especially recommended in order to view the developmental path of the form over time. Such study can illuminate the origin of the concept (whether embodied, cultural, or both), historical changes in structure and meaning, and the use of the form in situational contexts. The data obtained will serve to deepen understanding of conceptualization and the applicability of

the continuum concept to conceptual metaphor theory and to the general theoretical principles of cognitive linguistics.

References

Ackerknecht, Erwin H. 1982. *A short history of medicine.* Rev. edn. Baltimore MD: Johns Hopkins University Press.

Allan, Kathryn. 2006. On groutnolls and nog-heads: A case study of the interaction between culture and cognition in intelligence metaphors. In *Corpus-based approaches to metaphor and metonymy*, A. Stefanowitsch & S.T. Gries (eds.), 175–189. Berlin: Mouton de Gruyter.

Barlow, Michael & Suzanne Kemmer (eds.), 2000. *Usage based models of language*. Stanford CA: CSLI.

Bybee, Joan L. 1988. The diachronic dimension in explanation. *Explaining language universals*, John A. Hawkins, (ed.), 350–379. Malden, MA: Blackwell.

Bybee, Joan L. 2001. *Phonology and language use* [Cambridge Studies in Linguistics 94]. Cambridge: Cambridge University Press.

Clausner, Timothy C. & William Croft. 1997. Productivity and schematicity in metaphor. *Cognitive Science* (21): 247–282.

Croft, William. Forthcoming. Toward a social cognitive linguistics. In *New directions in cognitive linguistics*, V. Evans & S. Pourcel (eds.), Amsterdam: John Benjamins.

Enfield, Nick J. 2002. Ethnosyntax: Introduction. In *Ethnosyntax: Explorations in grammar and culture*, Nick J. Enfield (ed.), 1–30. Oxford: Oxford University Press.

Geeraerts, Dirk & Stefan Grondelaers. 1995. Looking back at anger: Cultural traditions and metaphorical patterns. In *Language and the cognitive construal of the world,* J.R. Taylor & R.E. MacLaury (eds.), 153–179. Berlin: Mouton de Gruyter.

Gevaert, C. 2002. The evolution of the lexical and conceptual field of ANGER in Old and Middle English. In *A changing world of words: Studies in English historical lexicography, lexicology and semantics*, J.E. Diaz Vera (ed.), 275–299. Amsterdam: Rodopi.

Gibbs, Ray W., Jr. 1999. Taking metaphor out of our heads and putting it into the cultural world. In *Metaphor in cognitive linguistics: Selected papers from the fifth International Cognitive Linguistics Conference, Amsterdam, July, 1997*, R.W. Gibbs Jr. & G.J. Steen (eds.), 145–166. Amsterdam: John Benjamins.

Greenberg, Joseph H. 1957. *Essays in linguistics.* Chicago IL: The University of Chicago Press.

Hoeniger, F. David. 1992. *Medicine and Shakespeare in the English Renaissance*. Newark DE: University of Delaware Press.

Hunston, Susan. 2002. *Corpora in applied linguistics* [Cambridge Applied Linguistics]. Cambridge: Cambridge University Press.

Koivisto-Alanko, Päivi & Heli Tisssari. 2006. Sense and sensibility: Rational thought versus emotion in metaphorical language. In *Corpus-based approaches to metaphor and metonymy* [Trends in Linguistics 171], A. Stefanowitsch & S.T. Gries (eds.), 191–207. Berlin: Mouton de Gruyter.

Kövecses, Zoltán. 2005. *Metaphor and culture: Universality and variation*. Cambridge: Cambridge University Press.

Lakoff, George. 1987. *Women, fire, and dangerous things*. Chicago IL: The University of Chicago Press.

Lakoff, George & Mark Johnson. 1999. *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York NY: Basic Books.

Lakoff, George & Zoltán Kövecses. 1987. The cognitive model of anger inherent in American English. *Cultural models in language and thought,* Dorothy Holland and Naomi Quinn (eds.), 195–221. Cambridge: Cambridge University Press.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar,* Vol.1: *Theoretical prerequisites*. Stanford CA: Stanford University Press.

MacArthur, Fiona. 2005. The competent horseman in a horseless world: Observations on a conventional metaphor in Spanish and English. *Metaphor and Symbol* 20(1): 71–94.

Nutton, Vivian. 1993. Humoralism. In *Companion encyclopedia of the history of medicine,* W.F. Bynum & R. Porter (eds.), 281–291. London: Routledge.

Oxford English dictionary online. 2005. Oxford: Oxford University Press.

Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* [Cambridge Studies in Linguistics 54]. Cambridge: Cambridge University Press.

Schmitt, Norbert & Ronald Carter. 2004. Formulaic sequences in action: An introduction. In *Formulaic sequences. Acquisition, processing, and use* [Language Learning and Language Teaching 9], N. Schmitt (ed.), 1–22. Amsterdam: John Benjamins

Tomasello, M. 2003. *Constructing a language: A usage-based theory of language acquisition.* Cambridge MA: Harvard University Press.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

From 'remaining' to 'becoming' in Spanish

The role of prefabs in the development of the construction *quedar(se)* + ADJECTIVE

Damián Vergara Wilson
University of New Mexico

Abstract

This study is based on Bybee & Eddington (2006), a synchronic study of the exemplar clusters formed by adjectives in four Spanish verb + adjective combinations used as constructions to denote a change of state (*ponerse* + ADJ, *hacerse* + ADJ, *quedarse* + ADJ, *& volverse* + ADJ). One main goal of the current study is to employ the exemplar model in a diachronic setting. This investigation studies the development of exemplar clusters of adjectives in the expression of 'becoming' *quedar(se)* + ADJECTIVE in four periods: the 13th, 15th, 17th, and 19th centuries. This study provides evidence that, a.) prefabs serve as the central members of exemplar categories, b.) prefabs have longevity, c.) categories mutate over time by becoming more centralized, changing central members, or by expanding.

1. Introduction

Spanish has a variety of verb + adjective combinations that participate in constructions used to indicate a change of state. Using the exemplar model to account for the distribution of adjectives in expressions of 'becoming', Bybee & Eddington (2006) studied the four most frequent verbs + adjective constructions used in Modern Spanish change-of-state expressions with animate subjects: *quedarse* + ADJ, *ponerse* + ADJ, *hacerse* + ADJ, and *volverse* + ADJ[1]. Examining data from the 13th, 15th, 17th, & 19th centuries, this study looks at changes in the exemplar categories formed by the open 'adjective' slot of the construction *quedar(se)* + ADJ[2]. The exemplar model provides a usage-based account for the increased productivity of this change-of-state construction over time; eventually it became the verb + adjective combination with the highest token frequency in 20th century data (Bybee & Eddington 2006). This investigation also provides insight into the process by which the verb *quedar(se)* came to be used as a verb of 'becoming'.

Part of the challenge of this analysis is that the verb *quedar(se)*, when used with an adjective, can mean 'to remain' (as in Ex. 1, below) whereas in other contexts it denotes a change of state (Ex. 2).

(1) *Mucho me duele que se aprovechen tan poco los consejos que os doy, y, pues todavía quedáis tan fatigado, os ruego os vais delante de aquella imagen de Nuestra Señora, que está allí, y le supliquéis os remedie.*
'It hurts me greatly that you take such little advantage of the advice that I give you, and, yet still you remain so fatigued, I beg that you go before that image of Our Lady that is over there, and you plead that she remedy you.' (*Vida y virtudes del venerable varón …* Juan de Ávila, Muñoz, Luis. 17th c.; Davies 2006)

(2) *E en aquella primera noche delas bodas que el conde & la condessa durmieron queda ella preñada.*
'And in that first night of the weddings that the count and the countess slept (together) she becomes (gets) pregnant.' (*Gran conquista de Ultramar*, anon., 13th c.; Davies 2006)

---

1. These four verbs, listed in the abstract and the introduction, are used most commonly to mean the following: *quedarse* 'to remain, to stay (reflexive/pronominal)', *ponerse* 'to put (reflexive/pronominal)', *hacerse* 'to make (reflexive/pronominal)' and *volverse* 'to return, to turn around (reflexive/pronominal)'. However, used in a change-of-state construction, the mean 'to become'.

2. The usage of a reflexive pronoun with the construction *quedar(se)* + ADJ is variable in my data; sometimes it is used, sometimes not. In order to indicate this variability, I put the pronoun in parenthesis.

The coexistence of two very different meanings points to the fact that this verb (as well as the other three studied by Bybee & Eddington (2006)), when used with an adjective to express a change of state, is a form-meaning pair that must be analyzed beyond the meaning of the individual parts, especially in regards to the actual verb. In these expressions of 'becoming' the semantic features of the verbs do not play a significant role in the choice of adjectives used with them (Eddington 1999; Bybee & Eddington 2006; 324) and it is more relevant to "ask not so much what the verb means, but what the overall construction means" (328). This construction, consisting of verb + adjective pairs, indicates a sense of becoming. As the verb is the fixed element of the construction, this study focuses on the categorization of adjectives, the open slot in the construction.

By investigating the categorization of adjectives in the construction *quedar(se)* + ADJ with a human subject, this study provides evidence that: a.) Prefabs serve as the central members of exemplar categories. b.) Central members of exemplar categories may retain their status over many centuries. As prefabs serve as central members, this shows that prefabs have longevity. c.) Categories may mutate over time in a variety of ways. The changes in categories observed in the data show that categories can become more centralized around a member with high token frequency. This central status may also pass from one adjective to another during the course of time. The construction *quedar(se)* + ADJ shows a general pattern of category expansion as exemplar clusters of adjectives gain in type frequency. In some cases, expansion may result in new exemplar clusters being formed. As with *way*-construction (Israel 1996), the construction *quedar(se)* + ADJ gained in productivity through an ongoing process of analogical extension.

2.1 Previous research on Spanish verbs of 'becoming'

Given that four common verb + adjective combinations (*quedarse, volverse, hacerse* or *ponerse* + adjective) are used to express the relatively synonymous idea of a change of state with an animate subject, the question has arisen as to what the factors are that influence a speaker to choose one combination over another. Reacting partially to the work of four previous researchers (Crespo 1949; Coste & Redondo 1965; Fente 1970; & Eberenz 1985) who posited, in some cases, discrete criteria for choosing among various combinations, Eddington (1999) applied four of the most concrete factors from the literature to 1,283 tokens of the following change-of-state verbs: *llegar a ser*, *ponerse*, *volverse*, *quedarse*, *convertirse*, *transformarse*, and *hacerse*. The factors applied were "(1) whether the verb's predicate is nominal or adjectival, (2) whether the change is gradual or abrupt, (3) whether the change occurs passively or actively, and (4) whether the noun or adjective of the predicate is expressed with *ser* or *estar*" (23). He found that semantic features did not effectively determine the choice of a certain

verb + adjective combinations in a given situation and that, even though there were noticeable tendencies, there was a "great deal of overlap and encroachment on the uses of each verb" (33). In other words, verb + adjective combinations are not discrete categories and some adjectives can be used with more than one verb.

In a questionnaire study on change of state verbs, Eddington (2002) presented native speakers with the task of choosing a change-of-state verb from a list to be used in combination with a certain adjective. The questionnaire was designed to apply the same factors with the same verbs as the previous analysis (Eddington 1999). The newer study supported his earlier findings that there is overlap in the usage of adjectives with these expressions of 'becoming'.

Considering the previous investigations by Eddington, using a corpus of both spoken and written Spanish, Bybee & Eddington (2006) studied 423 tokens consisting of the adjectives used in combination with the following four verbs: *ponerse*, *volverse*, *quedarse*, and *hacerse*. All of their tokens had an animate subject. They observed that even though there were strong tendencies observable in the data, any perceived boundaries between the categories of the complements of these verbs displayed a graded category membership. In order to better understand the nature of these constructions, they determined that it would be best to study the usage of verbal complements with a model that would allow some members of categories to be more central and others to be more marginal. A usage-based exemplar approach was chosen because it is based on the speaker's experience with language and can account for strong tendencies of verb + adjective combinations while allowing for the observed overlap of adjective use with different verbs. In order to develop exemplar clusters, a native speaker participant analyzed cards with the different adjectives types written on them. She then arranged the cards into groups according to perceived semantic similarity; the closer the cards, the closer the meaning. This organization of adjectives served as the basis for the organization of Bybee & Eddinton's (2006) exemplar clusters. Her pattern of organization was supported by two other experiments conducted with native speakers in Spain: a multidimensional scaling experiment and an acceptability experiment. Among their main findings were that "novel instances of verb + adjective sequences are based on analogies to previous experience and not on rules that refer to abstract features" which supports an exemplar model of representation based on the speaker's usage experience.

2.2 The exemplar model

Exemplar theory, although originally a theory used by psychologists in order to model perception and categorization, is an adequate and revealing theory for use in studying linguistic representation and change because of the way that it treats the speaker's experience with language. In exemplar theory, "each category is represented in memory by a large cloud of remembered tokens of that category"

(Pierrehumbert 2001: 140). Therefore, as Bybee & Eddington (2006) observe, each individual token of linguistic experience is categorized and mapped onto an identical exemplar, if present, thereby strengthening its representation. Each exemplar is formed by individually experienced tokens and novel tokens are produced based on similarity to established tokens. The defining characteristics for categorization, it follows, are not defined by a subset of binary or discreet features (Chandler 2002; Bybee & Eddington 2006). In the input, an incoming probe is compared to existing exemplars, and then classified according to the most similar one; the result is that representation is strengthened. If there is no similar exemplar, it is analogically classified according to perceived similarity to other existing ones. Because it is a theory based on language use, a token with no previous exemplar with which to associate may either 'die-out' or, if reinforced, serve as the basis for forming a new exemplar. As Chandler states (2002: 96), "Exemplar-based models imply that categories and categorization arise spontaneously when a probe enters our working memory and evokes into activation those memories that share experiential features with the probe." Because exemplar theory deals with different dimensions of categorization, it is logical that several dimensions of categorization are relevant to linguistic categorization including, but not limited to, semantic, phonological, morphological, situational, and pragmatic levels (Pierrehumbert 2001).

A central concept in the formation of categories is that exemplars with high token frequency serve as central members of these categories and could be thought of as prototypical, especially since they tend to display most of the features common in other members (Bybee & Eddington 2006). Also, like prototype categorization, there is 'family resemblance' where marginal members may share characteristics with the central members but not necessarily with one another. Furthermore, there is graded membership where some members are more central and some are more marginal, but the boundaries are not discrete (Bybee & Eddington 2006; Lakoff 1987).

The organization of exemplars into clouds, or clusters, has an impact on production as well. Pierrehumbert (2001) states that although there may be deeper causes, such as social and stylistic factors, the probability that a specific exemplar will be selected is proportionate to its strength of representation, or token frequency. Based on the high strength of their mental representation, it is unlikely that the adjective in high-frequency exemplars, such as *quedar(se) solo*, would be used with any other verb as an expression of becoming. Furthermore, this is unlikely because the high-frequency of these verb + adjective combinations indicates that they qualify as prefabs (Bybee2006; 25). In the case of novel expansion, the process is somewhat different. As pointed out in Bybee & Eddington (2006), redundant or marginal features could serve as the basis for the novel expansion of a category (see Chandler 2002). Because of this, it is not necessary to predict which features are chosen since they are all represented. Taken together, however, conceptual clustering can give an idea of how subsequent uses of a particular construction will manifest over time.

The idea of construction grammar is also necessary to the study at hand, as it was to Bybee & Eddington (2006) in their study of verb + adjective expressions of 'becoming.' Goldberg (1995: 4) takes the stance that "constructions are taken to be the basic units of language" and they are highly routinized even though they are readily extended to new contexts in principled ways. The construction under consideration, *quedar(se)* + ADJ, as with many constructions, is a form-meaning pair that must be analyzed not by the meaning of the individual parts, but as a whole unit. This study will look at the exemplar categories formed by the open 'adjective' slot in the construction and examine how previous uses affect subsequent ones either by routinized use or analogical extension.

3. Data & methodology

The data comes from narratives or narrative-like Peninsular Spanish works (such as narrative letters, novels, plays, or epic poems) that were chosen from a variety of electronic and internet sources (listed in 'Data Sources' below). Grammars, dictionaries, and legal documents were specifically excluded because of their dissimilarity to narrative works. Texts were chosen from four centuries: the 1200's, 1400's, 1600's and 1800's. The date of each text was confirmed based on bibliographic information given with the text or, if unclear, confirmed in one of two medieval bibliographies: Phaulhaber et al. (1984) or, Phaulhaber et al. (2002). The goal was to find occurrences of the expression of 'becoming', *quedar(se)* + ADJ as it was authentically used in the writing of each time period. Alternating centuries were chosen rather than consecutive ones in order to show as long of a time span as possible within the constraints of the present investigation.

The database generated for the analysis is based on entire texts that were loaded into a concordance program (*ConcApp*; Greaves 1993–2003) with the goal of extracting, at least 150 tokens of the construction *quedar(se)* + ADJ for each century. However, in the 1200's, only a total of 12 tokens were found. Word counts for each of the texts used were also entered into the database. One of the motivations was to be able to determine the true overall frequency for the construction in each century. By choosing to measure this in entire texts, it provides a more authentic measurement of frequency instead of extracting tokens only from sections of text in which the construction occurred with high frequency. All occurrences of *quedar(se)* + ADJ were analyzed individually to determine if they denoted a change of state (as opposed to meaning 'to remain') and to determine whether or not it occurred with a human subject. Animal subjects were accepted only if they were obviously personified (such as being able to speak and interact with humans).

Because the organization of exemplar clusters proposed in Bybee & Eddington (2006) was done by a native speaker, and was supported by two subsequent

experiments, I based the organization of my exemplar clusters on the ones proposed in their study. Also, when necessary, I used my own native-like intuition in the organization process. The idea is to demonstrate a theoretical mental representation and categorization of adjectives for Peninsular Spanish in four centuries. The figures that follow in this investigation provide a visual mapping of this theoretical mental organization of adjectives. Adjectives that are deemed similar are grouped into the same 'bubble', which is representative of an exemplar cluster. Central members of the clusters are the ones with the highest token frequency in the data and have a larger font, meant to roughly represent the number of tokens and the strength of representation (the higher the token frequency, the stronger the mental representation). As in Bybee (2006), central members with high token frequency are considered to be prefabs. The exemplar 'bubbles' are organized in relationship to one another by perceived similarity between the exemplar clusters; the more related to one another, the more overlap. The number in parenthesis reflects the token frequency of each adjective found in the data. In order to save space, if the translation of an adjective was given in a previous figure, it will not be given in following ones.

The current study follows the progress of two such clusters found in the data in order to demonstrate different ways in which exemplar clusters can develop: the *solo* cluster in four centuries (1200's, 1400's, 1600's & 1800's) and the *confuso/suspenso* cluster in three (1400's, 1600's & 1800's). In the examples that follow, all of the original spelling has been maintained as found in the data sources even if it is not consistent with the modern orthographic rules of Spanish. However, in the tables and figures all adjectives are listed in their singular masculine form.

4. Results

One of the central observations in studying the expression of becoming *quedar(se)* + ADJ, is that it has become more productive over the centuries. Even though this study focuses on the adjective types used in the construction, it is important to note that as exemplar categories expanded, there was also a rise in overall frequency.

Table 1. Overview of type and overall frequency in 5 centuries

	***Quedar(se)* + ADJ; expression of 'becoming'**			
	# of Tokens	**# of ADJ Types**	**Token to Type ratio**	**# of Tokens per 10,000 words**
1200's	12	10	1.2	0.01
1400's	169	91	1.86	1.07
1600's	155	75	2.07	4.42
1800's	164	88	1.86	2.36
1900's	122*	109*	1.12	1.83

Table 1 provides an overview of the frequency of the expression of 'becoming' *quedar(se)* + ADJ in five centuries: the 1200's to the 1800's from this study and the 1900's from Bybee & Eddington (2006). The overall frequency per 10,000 words is calculated with respect to the total number of words from the texts used. One of the most significant findings, as far as frequency goes, is in the jump between the 1200's and 1400's. In the 1200's only 12 tokens were found in the numerous works represented in the O'Neil (1999) and Davies (2006) corpora, and the overall frequency was very low (0.01 times/10,000 words). Yet by the 1400's, the overall frequency had multiplied by one hundred times (1.07 times/10,000 words) and there were multiple tokens in each work consulted.

With an increase in overall frequency from the 1200's to the 1400's, there is also an increase in type frequency. In the 1600's, the overall frequency is at its highest in the data (4.42 times/10,000 words) yet this also correlates with a lower number of types than any of the other centuries. In other words, this century has the highest ratio of tokens per type (2.07). This is explained by the fact that many of the peripheral members had more than just one token in this century as opposed to many novel tokens with just one occurrence. More succintly put, there are less novel tokens overall in the 1600's.

4.1 Clusters centering on *quedar(se) solo* 'to be left alone'

The prefab *quedar(se) solo* 'to be left alone' is the one with the highest token frequency in Bybee & Eddington's (2006) data from the 1900's. The adjective *solo* is also the only one that occurs in the construction under investigation more than one time in the 1200's. The endurance of this verb + adjective combination demonstrates that prefabs have longevity and the ensuing diachronic development shows a pattern of coherent organization of exemplar clusters around a central member.

4.1.1 *Clusters centering on* quedar(se) solo *in the 1200's*

Table 2 provides a list of all the tokens of *quedar(se)* + ADJ found in the data in the 1200's. The token frequency found in the works consulted in this century is very low, with the construction *quedar(se) solo* being the only one that appears more than once. Although seemingly insignificant with only three occurrences, it may actually be an indicator of emerging productivity as it will end up being the central member with the highest token frequency (28 occurrences) in Bybee & Eddington's (2006) study of 20th century data.

Example 3 shows how *quedar(se)* is used with the adjective *solo* 'alone' to denote a change of state. This example typifies all of the 13th and some of the 15th century examples whereby the change of state is brought about by the movement of people away from the human subject of the construction. Viewed this way, if the count 'remains' he will find himself without his subjects, the pilgrims. By remaining he would undergo a change of state and be left alone; both meanings are present:

Table 2. *Quedar(se)* + ADJ in the 1200's

Adjective (10 types)	#
solo 'alone'	3
muerto 'dead'	1
pobre 'poor'	1
sano 'healthy'	1
seguro 'safe'	1
ahorrado 'spared'	1
desafiado 'challenged'	1
con honra 'with honor'	1
sin dubda 'without doubt'	1
preñada 'pregnant'	1
Total	12

remaining and becoming. It is possible that this ambiguity may have been one of the original paths that opened up the verb *quedar (se)* to the analogical extension necessary for it to be used in a change-of-state construction.

(3) *E el conde quando vio que de otra manera no podia ser sino como queria el comun delos romeros no quiso ay quedar solo & fa zia lo mejor & cogio sus tiendas & fue se empos delos otros.*
'And when the count saw that there could be no other way than that which the majority of the pilgrims to Rome wanted, (he) didn't want to be left alone and did his best and gathered his tents and went after the others.' (*Gran conquista de Ultramar*, anon., 13th c.; Davies 2005)

The adjective types in the 1200's do not show as high a degree of semantic relatedness to each other, as in following centuries, and appear more miscellaneous. Although all of the adjective types, with the exception of *ahorrado* 'spared', will also appear in the construction *quedar(se)* + ADJ in following centuries, there is no centralized structure that can be applied to this set of adjectives as can be applied to the ones appearing in upcoming time periods. Even so, two sets of exemplar clusters could be proposed for this century in Figures 1 and 2 based on very general semantic values.

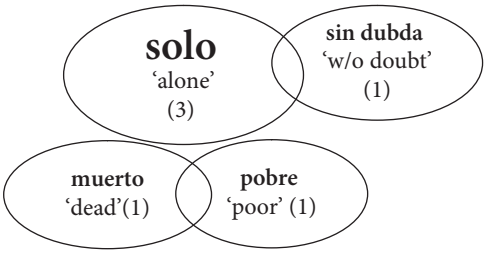


Figure 1. 1200's. Possible members of the *solo* clusters.

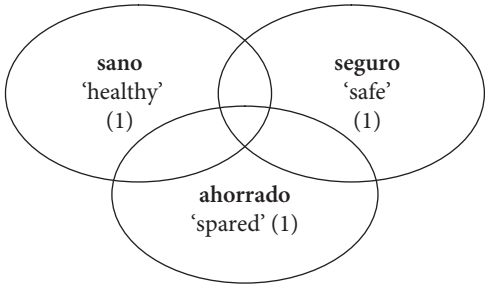


Figure 2. 1200's. The sano, seguro, ahorrado clusters.

In Figure 1, the exemplar cluster *quedar(se) sin dubda* 'to be left without (a) doubt' overlaps with the *solo* clusters as in later centuries; this overlap reflects a closer semantic relationship between these two clusters than the others in Figure 1. Here, the prepositional phrase has an adjectival function. This category will gain types over time and is referred to as the *quedar(se) sin* + NOUN 'left without + NOUN' clusters. It is similar to 'alone' because it indicates that the subject undergoes a change in which they are left without something that they previously had, be it human company or a possession (conceptual or physical).

The adjectives *pobre* 'poor', and *muerto* 'dead' relate to *solo* 'alone' because they are presented in the context of the data as undesirable states; in the examples, becoming poor or dead is presented as a negative change for the human subject. This is a very general theme that unites these exemplars and is not nearly as coherent as the tighter semantic relatedness observed in the organization of the *quedar(se) solo* exemplar clusters in following centuries. The opposite of this is undergoing a change that results in entering a desirable state. Figure 2 is based on this possible relationship between *sano* 'healthy', *seguro* 'safe' and, *ahorrado* 'spared'. This is an equally tenuous categorization as the one in Figure 1.

4.1.2 *Clusters centering on* quedar(se) solo *in the 1400's*

The adjectives in this century display more coherency in their organization around the central member *solo*; all of the other clusters convey the idea of being left without someone or something as seen in Figure 3.

Example 4 is similar to Example 3 in that it involves the movement of other people relative to the subject in order to bring about a change; by remaining inert, the subject is left alone. *Solo* is clearly the central member and, as opposed to the organization applied to the 1200's, the adjectives in this cluster have a strong, plausible semantic relationship with *solo.* One major difference between these two examples is that in Example 4, being left alone is portrayed as being a desirable outcome.

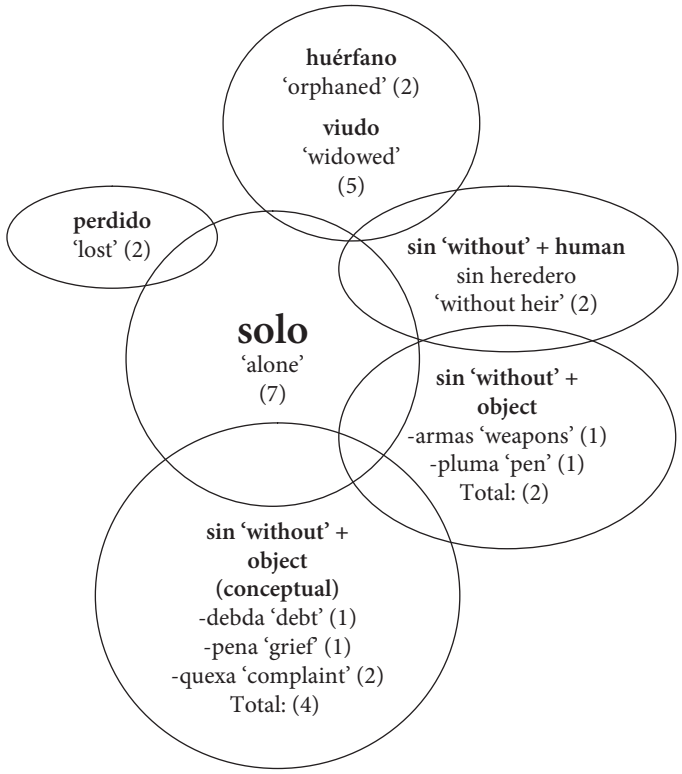


Figure 3. 1400's: Clusters centering on solo 'alone'.

(4) *E tornada en si la reyna mando que cada vno se retraxesse en su estancia & que quedasse sola enla camara.*
'And regaining consciousness the queen ordered that every one retreat from her room and that she be left alone in the quarters.' (*Oliveros de Castilla, Burgos*, anon., 1499; O'Neill 1999)

Another development in this century is in the types of nouns that can appear with the preposition *sin* 'without'. In order to emphasize possible subdivisions within this type, I divided it into *sin* + human, *sin* + object, and *sin* + conceptual object. It would also be reasonable to propose that these all fit into the same exemplar cluster. The important thing to note is that there are five types of nouns used in the construction *quedar(se) sin* + NOUN. Example 5 shows a similar change of state as the one in *quedar(se) solo*; it depends on the actions of others (the death of the lord and his daughter) in order to bring about the change. As with *quedar(se) solo,* the change brought about in the construction *quedar(se) sin* + NOUN is one in which the subject is left without something.

(5) *E visto que nos se ra grand daño sy asy perdemos el señor & la hija & quedamos syn herederos nos somos en grand congoja.*
'And seeing that great damage shall come about if we loose the lord and the daughter like that & we are left without heirs we are in great anguish.' (*Historia de la Linda Melosina*, anon. 15th c.; O'Neill 1999)

The clusters with *viudo* 'widowered' and *huérfano* 'orphaned' are related to *solo* because they both indicate that there was a change in which the subject was left without an immediate family member (or members). As demonstrated in example 6, this change depends on outside action, in this case the death of the queen.

(6) *Enla tierra de ansaj avia vn potente rrey al qual no avia quedado sy no vna hija la qual avia avi- do de su muger que enel ora del parto murio & quedo biudo mas el rrey hjzo criar la hija muy honorable mente.*
'In the land of Ansaj there was a powerful king to whom no one was left but a daughter, who he had had from his wife, who in the moment of birth died & (he) became widowered, but the king had the daughter raised honorably.' (*Historia de la Linda Melosina*, anon., 15th c.; O'Neill 1999)

The adjective *perdido* 'lost' is placed in the *solo* clusters because it expresses a change in which the subject finds themselves without a clear understanding of where they are in relationship to a previous trajectory (physical or conceptual). By getting lost, the subject is left without direction as in Example 7.

(7) *Vna muy bonita moça; avnque queda agora perdida la pecadora, porque tenía a Celestina por madre e a Sempronio por el principal de sus amigos.*
'A very pretty girl; even though the sinner is now lost, because she had Celestina for a mother and Sempronio as her main friend. (*La Celestina*, Rojas, 1499; BVMC)

The proposed organization of exemplar clusters centering *quedar(se) solo* in the 1400's is also justified by the fact that it shows a pattern of cluster organization and adjective types that are replicated in upcoming centuries. The clusters in the 1600's and the 1800's show a very similar structure of organization in which closely related adjective types overlap with the central member *solo* in order to represent their coherent semantic similarity.

4.1.3 *Clusters centering on* quedar(se) solo *in the 1600's*

As seen in Figure 4, the clusters from this century show the continuity in the organization of these exemplar clusters to the ones from the 1400's. Even though the proposed categorization of exemplars is theoretical, the continuity in the structure of this set of clusters shows perseverance; previous usage affects latter

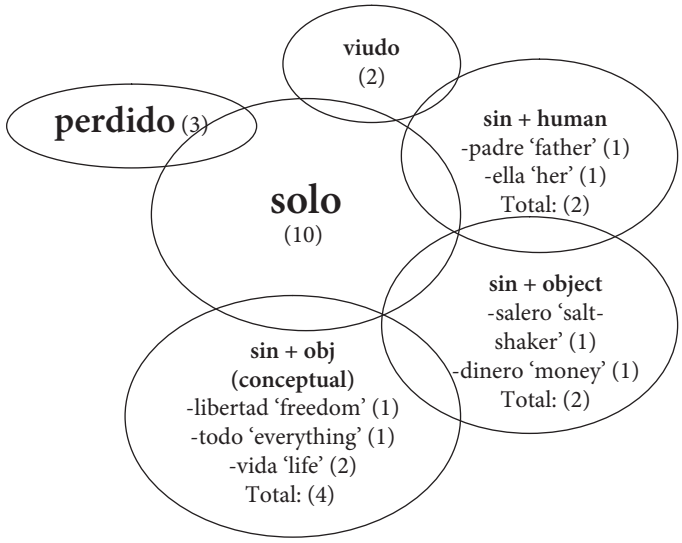


Figure 4. 1600's: Clusters centering on *solo*.

usage. There is an obvious central member (*solo*: 10 tokens) and a clear semantic similarity between adjectives that center on the concept of the subject being left without something.

Examples 8, and 9 show usages of *quedar(se)* + ADJ with two adjectives. In Example 8, the 3s conjugation of *quedar* activates two complements: the dame who was left satisfied, and Don Diego who was lost. Example 9 is similar to 8 in that the 3s verb is used to activate two subjects (the second time by implication) but the adjectives (prepositional phrases in this case) are both manifestations of *quedar(se) sin* + NOUN and both belong in the *solo* clusters.

(8) *Y desterrada por seis años de la ciudad, no declarándose más el caso por la opinión de doña Inés, con que la dama quedó satisfecha en parte, y don Diego más perdido que antes*.
'And banished for six years from the city, not testifying more in the case for the Doña Inés' opinion, by which the dame was left satisfied in part, and Don Diego more lost than before. (*La inocencia castigada & El jardín engañoso, Sotomayor*, 17th c.; BVMC)

(9) *Para dar a los que piden de beber la colación; con que tu padre se queda sin salero, (y) tú, señor sin padre*.
'In order to give to those who ask to drink the collation; because of which your father will be left without a saltshaker, (and) you, sir without a father. (*Abre el ojo*, Rojas Zorrilla, 17th c.; BVMC)

4.1.4 *Clusters centering on* quedar(se) solo *in the 1800's*

The organization of exemplar clusters follows the same patterns of the previous two centuries studied; this set of exemplar clusters has maintained enough consistency through time to show only minor changes since the 1400's. There is some evidence that *solo* is a central member in the 1200's. This demonstrates how repeated patterns in formulaic language, as with the categorization of adjectives shown here, may endure for many centuries. One of the advantages in choosing every other century, as opposed to consecutive ones, is that it covers a longer span of time. Considered this way, the remarkable similarity among the exemplar clusters from the 1400's to the 1800's represents a range of five centuries in which this notable similarity persists.

One of the new additions is of opposites as with the adjectives *unido* 'united', *reunido* 'reunited', and *convidado* 'invited' shown in Figure 5 and in Examples 10 (*unido*) and 11 (*convidado*). Bybee & Eddington also have opposites in their data and observe that they actually share many features 'while having a negative value for one important feature' (2006; 332). One of the advantages of using the exemplar model is that a set of features does not need to match exactly to another set in order to provide a basis from which to produce novel uses. Like *solo*, the adjectives *reunido* and *convidado* used in the construction *quedar(se)* +ADJ show a change in

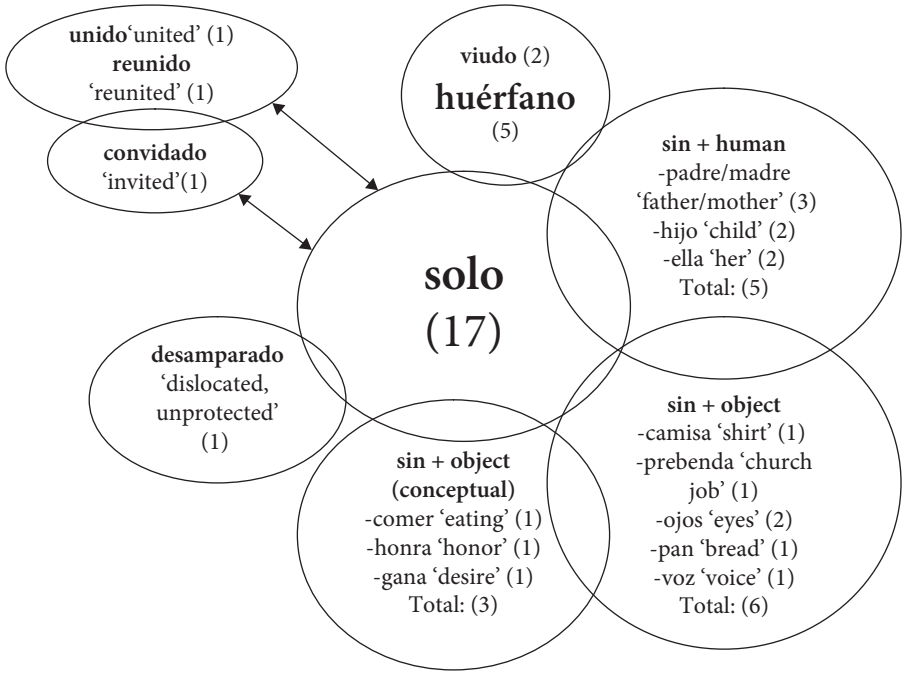


Figure 5. 1800's: Clusters centering on solo.

an individual's relationship to a group. Only here, *reunido* and *convidado* indicate the unification of human subjects whereas *solo* attests to the division of a group of human subjects.

(10) *Me vio, le vi, y nuestras almas quedaron unidas para siempre.*
'She saw me, I saw her, and our souls were united forever.
(*Angela*, Tamayo, 19th c.; AHCT)

(11) *Conque, con Dios, don Plácido; queda usted convidado para la boda y para el bautizo.*
'Because of which, with God, Don Plácido; you are invited to the wedding and the baptism.' (*A fuerza de arrastrarse*, Echegaray, 19th c.; AHCT)

The use of two adjectives with the single verb *quedar(se)* in Example 12 demonstrates that these two adjectives, *huerfano* 'orphaned', and *desamparado* 'dislocated, separated', are perceived by the speaker (or writer in this case) as being similar and adequate for usage in the construction. Used together, the adjectives strengthen the argument that the subject is on a path toward being left orphaned and left without ennobling and aggrandizing ideals. Since *desamparado* is also a novel usage in the *quedar(se) solo* clusters, this example could be a demonstration of the exemplar model at work whereby a novel usage is created based on its similarity to established usages, a process also studied in Bybee & Eddington (2006). In Example 12, the writer is addressing the people of his home village as a whole (*pueblo*: people) and describing a situation in which they will be metaphorically orphaned and separated from a set of ideals. Since *huerfano* has been used consistently in the *quedar(se) solo* clusters since the 1400's, it would be a readily accessible adjective to use in the open slot of the construction *quedar(se)* + ADJ.

(12) *No por eso ¡oh pueblo de las grandes melancolías! quedarás huérfano y desamparado de ideales que te sublimen y ennoblezcan.*
'Not for this, oh people of the great melancholies, shall you become orphaned and dislocated (separated) from ideals that may aggrandize and ennoble you.'
(*Al primer vuelo*, Pereda, 1896; BVMC)

The similarity between the exemplar clusters centering on *quedar(se) solo* in the data from the 15th to the 19th century and Bybee & Eddington's (2006) data from the 20th century indicates some relevant factors. First, the diachronic process demonstrated in this study reconciles plausibly with the 20th century (Modern Spanish) data and findings in Bybee & Eddington (2006); in their study, out of a total of 122 tokens of *quedar(se)* + ADJ, 28 were with *solo*. Table 2 is taken

from page 330 of their investigation and shows all of the adjectives in their clusters centering on *quedar(se) solo* (I collapsed their two columns for spoken and written data into one column).

Table 3. Adjectives related to *solo* used with *quedarse* (Bybee & Eddington 2006; 332)

Adjective	***# (written & spoken)***
solo 'alone'	28
soltera 'single, unmarried'	3
aislado 'isolated'	2
a solas 'alone'	1
sin novia 'without a girlfriend'	1
OPPOSITE	
emparejado 'paired with'	1

Because the prefab *quedar(se) solo* demonstrates such an increase in productivity, more than any other adjective, it is more evidence that the 3 tokens in the 1200's do point to the emerging significance of the prefab. Accordingly, the diachronic development of the exemplar clusters centering on *solo* shows that prefabs have longevity; previous usage effects latter usage. It is highly relevant that the only notable change in the organization of exemplar clusters occurred between the 1200's and the 1400's. In the 1200's, the construction *quedar(se)* + ADJ was very infrequent and the exemplar clusters proposed weren't very cohesive. Conceivably this is where the construction itself was emerging in written texts. It is possible that *quedar(se) solo* may have been a factor in the gradual re-analysis of *quedar(se)* as a verb that came to participate in a change-of-state construction. It shows a path from where it could have been used earlier to mean 'to remain', and evolved into a construction that could include the two meanings 'to remain' and 'to become' (Examples 3 & 4), to later be used as an unambiguous change-of-state expression. Regardless, from the 1200's to the 1400's, this set of exemplar clusters went from being difficult to organize into a reasoned set of clusters to developing a centralized, semantically coherent structure based on the adjective *solo* that perseveres into modern usage.

4.2 Clusters centering on quedar(se) confuso 'to become confused/suspenso 'astonished'

This set of clusters shows a different pattern of diachronic development than the clusters centering on *quedar(se) solo*. One of the differences is that there is no inherent semantic ambiguity in the usage of the construction *quedar(se)* + ADJ in the *confuso/suspenso* clusters; they are all clearly expressions of 'becoming' and do not implicate any kind of 'remaining'. Perhaps this is due, in part, to the fact that these

clusters do not appear until the 15th century in this study. By this time, the construction *quedar(se)* + ADJ had a much higher frequency of usage (1.07/10,000 words) in contexts where it was used to express a change of state than in the 13th century (0.01/10,000 words). Also, the clusters centering on *quedar(se) solo* demonstrate that at least some of the adjectives used in the construction supported a coherent organization exemplar clusters focused on a central adjective and a central concept.

4.2.1 *Clusters centering on* quedar(se) confuso/suspenso *in the 1400's*
As with the proposed clusters centering on *solo* in the 1200's, there is no strong centrality to this cluster. Only three types could be posited for this century with their similarity based on the idea of entering a psychological state in which one lacks mental clarity as in Examples 13 and 14. While *confuso* 'confused' is present in this set of clusters, *suspenso* 'astonished' is not (Figure 6).

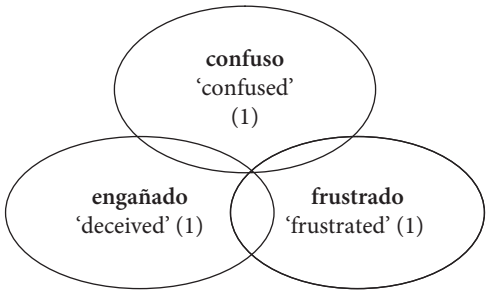


Figure 6. 1400's: Clusters centering on *confuso.*

(13) *sylo entendemos asi esforçado que seneca quede confusso en su dezir commo el aya escripto en diversos lugares.*
'If we understand it, by great pains, that Seneca may be confused in saying how he may have written in various places.' (*Defensa de virtuossas mugeres*, anon., 15th c.; O'Neill 1999)

(14) *Quien non faze el primero Pora quedar enganyado Quien faze lo postrimero Mucho deue ser culpado.*
'He who does not do it the first time is because of being deceived. He who does not do it the last time should himself to blame.' (*Cancionero castellano y catalán de París*, anon., 15th c.; O'Neill 1999)

4.2.2 *Clusters centering on* quedar(se) confuso/suspenso *in the 1600's*
This set of clusters has expanded in order to include a variety of adjective types and the two adjectives *suspenso* 'astonished' and *confuso* 'confused' appear to express the essential idea in this exemplar cluster. In Example 15, they are used together with

the same subject. It is possible that instead of being based on just one or two exemplars, this set of clusters is based more firmly on the idea of reacting to a sudden, unexpected situation that may leave the subject astonished or confused. Although the set of clusters in Figure 7 shows that the construction *quedar(se)* + ADJ has become much more productive than it had been in the 1400's, it does not lend itself to the highly centralized organization that *quedar(se) solo* does where there is a clearly prevailing central member. However, the double usage of *suspenso* and *confuso* reinforces the idea that these concepts are similar in the mind of the writer and may justify their membership in the same set of exemplar clusters. Used together, they reinforce an idea of a subject who reacted to an unexpected situation, the discovery of some papers, and came out surprised and confused as a result.

(15) *Suspenso y confuso quedé, no sabiendo quién pudiese ser el dueño de aquellos papeles.*
'Surprised and confused I became, not knowing who could possibly be the owner of those papers.' (*La fantasma de Valencia*, Castillo Solórzano, 1652; BVMC)

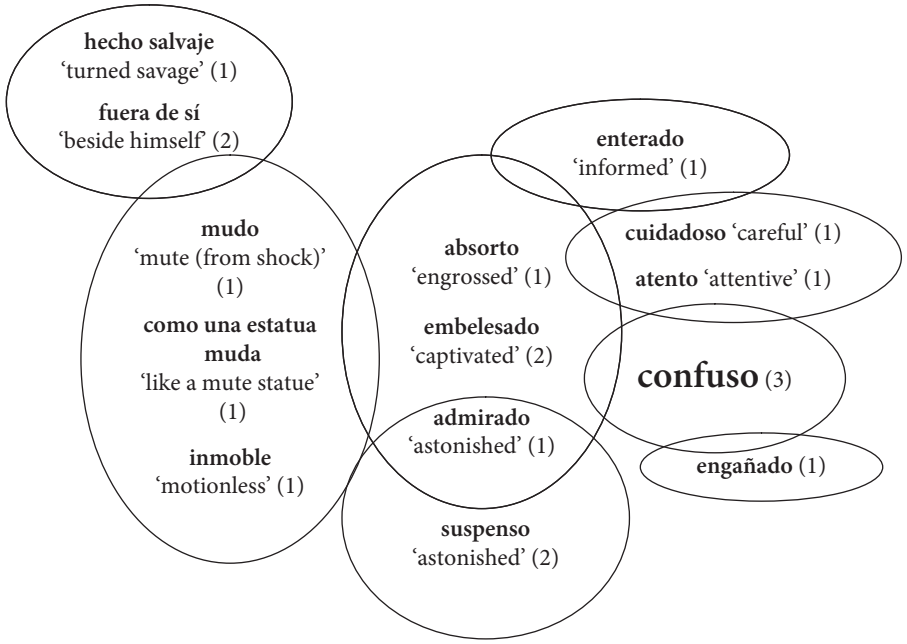


Figure 7. 1600's: Clusters centering on *confuso*.

By this century, this set of clusters now includes the idea of being surprised, and the resulting physical or psychological state (i.e., *fuera de sí* 'beside oneself' as

in Example 16, *mudo* 'mute', or *hecho salvaje* 'turned savage'). Another possibility is that one reacts to a surprising or perplexing situation by becoming careful (*cuidadoso*) or attentive (*atento*). Being deceived (*engañado*) is another possible result of being confused and shows some continuity with the 1400's.

(16) *Se entró, y sin hablar palabra, ni mirar en nada, se puso dentro de la cama donde estaba don Diego, que viendo un caso tan maravilloso, quedó fuera de sí.*
'She entered, and without saying a word, nor looking at anything, she got in bed where Don Diego was, who upon seeing such a marvelous case, was left beside himself'. (*La inocencia castigada & El jardín engañoso*, Sotomayor, 17th c.; BVMC)

4.2.3 *Clusters centering on* quedar(se) confuso/suspenso *in the 1800's*

In this century I propose that *suspenso* 'astonished' is the central member; even though it occurs only four times (Figure 8), it has the highest token frequency of

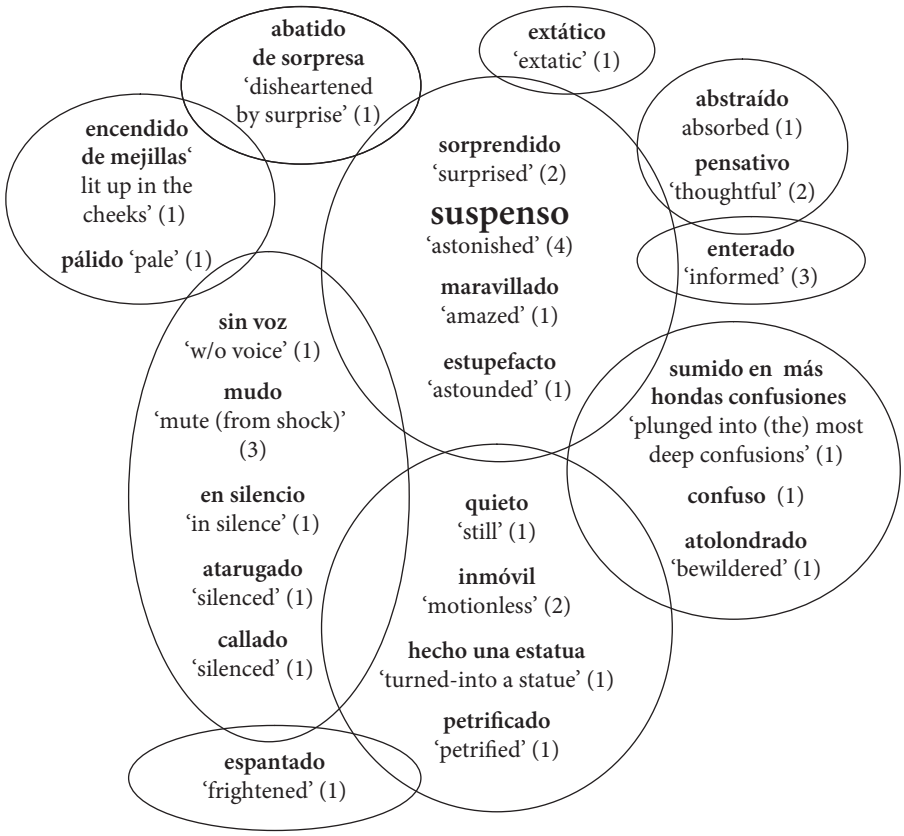


Figure 8. 1800's Clusters centering on *suspenso.*

any other adjective. In a previous, similar study (Wilson 2005) there was strong evidence that the central member in the 1800's was *sorprendido* 'surprised' as it appeared 26 times in the data. However, this may be attributable to the data used. Wilson (2005) used Davies' *Corpus del Español* (in this case accessed in 2005) which does not have a filter for sorting data according to region and would include many texts from Latin America. As all of the data from the present study comes from Peninsular Spanish, the results could point to regional differences or it could be simply a case of diachronic change. In any case, the clusters centering on *suspenso* 'astonished' display a much more coherent structure in the organization of exemplar clusters, meaning that there is a more commonsensical semantic similarity between the adjectives found in the data.

The focus has shifted from *confuso* 'confused' to *suspenso* 'astonished', showing evidence that even though the one central member may have gained strength in usage whereas the other has begun to lag, the idea remains the same. In the previous study (Wilson 2005) the equivalent set of clusters changed central members from *admirado* 'amazed' as the central member (26 occurrences) in the 1600's to *sorprendido* in the 1800's. In Bybee & Eddington's (2006) 20th century data, the central member was also *sorprendido* with 7 total tokens. It must be noted that their data included both Latin American and Spanish sources, and both written and spoken texts. However, it is revealing that in my own two diachronic studies, even though the data was different, there was a pattern of shift regarding the central members but the general idea remained the same.

The idea of becoming motionless as a reaction to an unexpected situation has become more productive as in Examples 18–20. Example 19 is another demonstration of how two adjectives appear with the verb in order to create a more vivid image. Here the subject reacts to the jokes of the scribe by becoming silenced and lit up in the cheeks. As with the previous double adjectives, this bolsters the already comprehensible semantic similarities that lead these two adjectives to be grouped into the same set of clusters.

(18) *Parecía que se había quedado mudo o que no sabía qué decir.*
'It seemed like (he) had become mute or that (he) didn't know what to say.'
(*Al primer vuelo*, Pereda, 1896; BVMC)

(19) *Con esta pregunta se quedó Leto bastante atarugado y algo encendido de mejillas: ¡le había dado tantas bromas el fiscal con la Escribana mayor!*
'With that question, Leto was left silenced and somewhat lit up in the cheeks: The agent with the main scribe had played so many jokes on him! ' (*Al primer vuelo*, Pereda, 1896; BVMC)

(20) *Valentín, que se paró, se quedó inmóvil de súbito, como si se hubiera convertido en piedra.*
'Valentín, who stopped, suddenly became motionless, as though he had turned to stone.' (*Al primer vuelo*, Pereda, 1896; BVMC)

One of the things that this set of clusters shows is evidence of continuing category expansion. Based on the central idea of reacting to an unexpected situation that causes the subject to enter a surprised or confused mental state, these clusters demonstrate expansion into resulting physical states observable by the 1600's. Israel (1996) attributes a gain in productivity to analogical extension. In this case, a set of clusters gains new clusters through analogical extension, and the category is expanded.

Another observation made in Israel (1996) was that the open slot in the *way*-construction demonstrated schematization effects whereby new usages were based on a set of established usages instead of just one central type. The aggregate of old usages created a general idea from which to generate new ones. It is possible that the general idea of being surprised and/or confused is the driving force behind the expansion of this category instead of the actual types or tokens. This would contribute to an explanation for the lack of emergence in these data of a clear central member with high token frequency. Another factor is that there are more synonyms for *suspenso* than there are for *solo*. This could contribute to the immediate accessibility of exemplars by providing the language user with a larger set of words with the same meaning to insert into the open slot of the construction. More data and research is needed to develop these ideas.

5. Conclusions

The exemplar model applied to construction grammar reveals overarching tendencies in the development of constructions and provides a model by which a construction gains in productivity. As in the case of *quedar(se) solo*, central members of exemplar categories may retain their status as central members for many centuries showing that prefabs have longevity. Clusters will mutate over time, some in different manners than others. They may become more centralized, as demonstrated with the clusters centered on *quedar(se) solo*; the central member gains in token frequency and it is more plausible to propose a highly organized set of exemplar clusters based on the central member. Even the clusters centering on *quedar(se) confuso/suspenso* show a centralization effect; the category conveys the

same overall concept over time and the tokens are more plausibly organized into a set of exemplar clusters by the 1800's. However, these clusters show a different pattern of development than the *quedar(se) solo* clusters. The central member appears to shift from *confuso* to *suspenso*, in these data, and later to *sorprendido* in Bybee & Eddington's (2006) data. As with the *way*-construction studied by Israel (1996), the construction *quedar(se)* + ADJ shows a rise in productivity; an established set of clusters gains new clusters through analogical extension and the category is expanded as a result (Israel 1996).

Data Sources

Association for Hispanic Classical Theater, INC. (AHCT). http://www.comedias.org/. Accessed in Spring 2007.

Davies, Mark. 2006. Corpus del español. www.corpusdelespanol.org. Accessed in Fall 2006.

Biblioteca Virtual Miguel de Cervantes (BVMC). http://www.cervantesvirtual.com. Accessed in Spring 2007.

LEMIR Revista Electrónica sobre Literatura Española Medieval y Renacimiento. http://parnaseo.uv.es/Lemir.htm. University of Valencia. Accessed in Spring 2007.

O'Neil, John. 1999. *Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings,* CD-ROM. Madison and New York; Hispanic Seminary of Medieval Studies.

References

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 529–551.

Bybee, Joan L. & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2): 323–355.

Chandler, Steve. 2002. Skousen's analogical approach as an exemplar-based model of categorization. In *Analogical modeling: An exemplar-based approach to language,* R. Skousen, D. Lonsdale & D.B. Parkinson (Eds), 51–105. Amsterdam: John Benjamins.

Coste, Jean & Augustín Redondo. 1965. *Syntaxe de l'espagnol moderne*. Paris: Société d'Edition d'Enseignement Superieur.

Crespo, Luis A. 1949. To become. *Hispania* 32: 210–12

Diccionario de la Real Academia Española (DRAE). 2006. www.rae.es. (Accessed in October 2006)

Eberenz, Rolf. 1985 Aproximación structural a los verbos de cambio en Iberorromance. In *Linguistique comparée et typologie des langues romanes*, Jean-Claude Bouvier (Ed.), 460–75. Aix en Provence: Université de Provence.

Eddington, David. 1999. On 'becoming' in Spanish: A corpus analysis of verbs expressing change of state. *Southwest Journal of Linguistics* 18: 23–46.

Eddington, David. 2002. Disambiguating Spanish change of state verbs. *Hispania* 85: 921–29

Fente, R. 1970. Sobre los verbos de cambio o 'devenir'. *Filología Moderna* 38: 157–72
Israel, Michael. 1996. The way constructions grow. In *Conceptual structure, discourse and language,* A. Goldberg (Ed.), 217–231. Stanford CA: CSLI.
Goldberg, Adele. 1995. *Constructions. A construction grammar approach to argument structure*. Chicago IL: The University of Chicago Press.
Greaves, Chris. 2005. Concapp Windows application V4. http://www.edict.com.hk/PUB/concapp/. (Downloaded in Spring 2007)
Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago IL: The University of Chicago Press.
Phaulhaber et. al. 2002. Bibliografía española de textos antiguos. Vol. 2002, Number 2 (July). http://sunsite.berkeley.edu/Philobiblon/phhmbe.html. (Accessed in Spring 2007)
Phaulhaber et. al. 1984. *Bibliography of old Spanish texts*, 3rd Edn. Madison WI: Hispanic Seminary of Medieval Studies.
Pierrehumbert, J. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency and the emergence of linguistic structure,* J. Bybee & P. Hopper (Eds), 137–157. Amsterdam: Benjamins.
Wilson, Damián. 2005. '*Quedarse* + ADJECTIVE since the 1200's: An exemplar approach'. Final project for graduate seminar Frequency Effects and Emergent Grammar, Prof. J.L. Bybee, UNM, Spring 2005.

Author index

Note: page numbers 297–638 refer to Volume 2.

Subject index

Note: page numbers 297–638 refer to Volume 2.

Typological Studies in Language

A complete list of titles in this series can be found on the publishers' website, *www.benjamins.com*

87 **CYFFER, Norbert, Erwin EBERMANN and Georg ZIEGELMEYER (eds.):** Negation Patterns in West African Languages and Beyond. *Expected August 2009*

86 **MAHIEU, Marc-Antoine and Nicole TERSIS (eds.):** Variations on Polysynthesis. The Eskaleut languages. 2009. ix, 312 pp.

85 **GIVÓN, T. and Masayoshi SHIBATANI (eds.):** Syntactic Complexity. Diachrony, acquisition, neuro-cognition, evolution. 2009. vi, 553 pp.

84 **NEWMAN, John (ed.):** The Linguistics of Eating and Drinking. 2009. xii, 280 pp.

83 **CORRIGAN, Roberta, Edith A. MORAVCSIK, Hamid OUALI and Kathleen M. WHEATLEY (eds.):** Formulaic Language. Volume 2. Acquisition, loss, psychological reality, and functional explanations. 2009. xxiv, 361 pp.

82 **CORRIGAN, Roberta, Edith A. MORAVCSIK, Hamid OUALI and Kathleen M. WHEATLEY (eds.):** Formulaic Language. Volume 1. Distribution and historical change. xxiv, 315 pp.

81 **CORBETT, Greville G. and Michael NOONAN (eds.):** Case and Grammatical Relations. Studies in honor of Bernard Comrie. 2008. ix, 290 pp.

80 **LAURY, Ritva (ed.):** Crosslinguistic Studies of Clause Combining. The multifunctionality of conjunctions. 2008. xiv, 253 pp.

79 **ABRAHAM, Werner and Elisabeth LEISS (eds.):** Modality–Aspect Interfaces. Implications and typological solutions. 2008. xxiv, 422 pp.

78 **HARRISON, K. David, David S. ROOD and Arienne DWYER (eds.):** Lessons from Documented Endangered Languages. 2008. vi, 375 pp.

77 **SEOANE, Elena and María José LÓPEZ-COUSO (eds.):** Theoretical and Empirical Issues in Grammaticalization. In collaboration with Teresa Fanego. 2008. x, 367 pp.

76 **LÓPEZ-COUSO, María José and Elena SEOANE (eds.):** Rethinking Grammaticalization. New perspectives. In collaboration with Teresa Fanego. 2008. x, 355 pp.

75 **FRAJZYNGIER, Zygmunt and Erin SHAY (eds.):** Interaction of Morphology and Syntax. Case studies in Afroasiatic. 2008. v, 234 pp.

74 **KURZON, Dennis and Silvia ADLER (eds.):** Adpositions. Pragmatic, semantic and syntactic perspectives. 2008. viii, 307 pp.

73 **ANSALDO, Umberto, Stephen MATTHEWS and Lisa LIM (eds.):** Deconstructing Creole. 2007. xii, 292 pp.

72 **NÆSS, Åshild:** Prototypical Transitivity. 2007. x, 240 pp.

71 **NEDJALKOV, Vladimir P. (ed.):** Reciprocal Constructions. With the assistance of Emma Š. Geniušienė and Zlatka Guentchéva. 2007. xxiii, 2219 pp. (5 vols.).

70 **ZÚÑIGA, Fernando:** Deixis and Alignment. Inverse systems in indigenous languages of the Americas. 2006. xii, 309 pp.

69 **ARANOVICH, Raúl (ed.):** Split Auxiliary Systems. A cross-linguistic perspective. 2007. vii, 277 pp.

68 **ABRAHAM, Werner and Larisa LEISIÖ (eds.):** Passivization and Typology. Form and function. 2006. x, 553 pp.

67 **VESELINOVA, Ljuba N.:** Suppletion in Verb Paradigms. Bits and pieces of the puzzle. 2006. xviii, 236 pp.

66 **HICKMANN, Maya and Stéphane ROBERT (eds.):** Space in Languages. Linguistic Systems and Cognitive Categories. 2006. x, 362 pp.

65 **TSUNODA, Tasaku and Taro KAGEYAMA (eds.):** Voice and Grammatical Relations. In Honor of Masayoshi Shibatani. 2006. xviii, 342 pp.

64 **VOELTZ, F. K. Erhard (ed.):** Studies in African Linguistic Typology. 2006. xiv, 426 pp.

63 **FILIMONOVA, Elena (ed.):** Clusivity. Typology and case studies of the inclusive–exclusive distinction. 2005. xii, 436 pp.

62 **COUPER-KUHLEN, Elizabeth and Cecilia E. FORD (eds.):** Sound Patterns in Interaction. Cross-linguistic studies from conversation. 2004. viii, 406 pp.

61 **BHASKARARAO, Peri and Karumuri Venkata SUBBARAO (eds.):** Non-nominative Subjects. Volume 2. 2004. xii, 319 pp.

60 **BHASKARARAO, Peri and Karumuri Venkata SUBBARAO (eds.):** Non-nominative Subjects. Volume 1. 2004. xii, 325 pp.

59 **FISCHER, Olga, Muriel NORDE and Harry PERRIDON (eds.):** Up and down the Cline – The Nature of Grammaticalization. 2004. viii, 406 pp.

58 **HASPELMATH, Martin (ed.):** Coordinating Constructions. 2004. xcv, 578 pp.

57 **MATTISSEN, Johanna:** Dependent-Head Synthesis in Nivkh. A contribution to a typology of polysynthesis. 2003. x, 350 pp.

56 **SHAY, Erin and Uwe SEIBERT (eds.):** Motion, Direction and Location in Languages. In honor of Zygmunt Frajzyngier. 2003. xvi, 305 pp.

55 **FRAJZYNGIER, Zygmunt and Erin SHAY:** Explaining Language Structure through Systems Interaction. 2003. xviii, 309 pp.

54 **AIKHENVALD, Alexandra Y. and R.M.W. DIXON (eds.):** Studies in Evidentiality. 2003. xiv, 349 pp.

53 **GIVÓN, T. and Bertram F. MALLE (eds.):** The Evolution of Language out of Pre-language. 2002. x, 394 pp.

52 **GÜLDEMANN, Tom and Manfred von RONCADOR (eds.):** Reported Discourse. A meeting ground for different linguistic domains. 2002. xii, 425 pp.

51 **NEWMAN, John (ed.):** The Linguistics of Sitting, Standing and Lying. 2002. xii, 409 pp.

50 **FEIGENBAUM, Susanne and Dennis KURZON (eds.):** Prepositions in their Syntactic, Semantic and Pragmatic Context. 2002. vi, 304 pp.

49 **WISCHER, Ilse and Gabriele DIEWALD (eds.):** New Reflections on Grammaticalization. 2002. xiv, 437 pp.

48 **SHIBATANI, Masayoshi (ed.):** The Grammar of Causation and Interpersonal Manipulation. 2002. xviii, 551 pp.

47 **BARON, Irène, Michael HERSLUND and Finn SØRENSEN (eds.):** Dimensions of Possession. 2001. vi, 337 pp.

46 **AIKHENVALD, Alexandra Y., R.M.W. DIXON and Masayuki ONISHI (eds.):** Non-canonical Marking of Subjects and Objects. 2001. xii, 364 pp.

45 **BYBEE, Joan and Paul J. HOPPER (eds.):** Frequency and the Emergence of Linguistic Structure. 2001. vii, 492 pp.

44 **VOELTZ, F. K. Erhard and Christa KILIAN-HATZ (eds.):** Ideophones. 2001. x, 436 pp.

43 **GILDEA, Spike (ed.):** Reconstructing Grammar. Comparative Linguistics and Grammaticalization. 2000. xiv, 269 pp.

42 **DIESSEL, Holger:** Demonstratives. Form, function and grammaticalization. 1999. xii, 205 pp.

41 **FRAJZYNGIER, Zygmunt and Traci S. WALKER-CURL (eds.):** Reciprocals. Forms and functions. Volume 2. 2000. xii, 201 pp.

40 **FRAJZYNGIER, Zygmunt and Traci S. WALKER-CURL (eds.):** Reflexives. Forms and functions. Volume 1. 2000. xiv, 286 pp.

39 **PAYNE, Doris L. and Immanuel BARSHI (eds.):** External Possession. 1999. ix, 573 pp.

38 **SIEWIERSKA, Anna and Jae Jung SONG (eds.):** Case, Typology and Grammar. In honor of Barry J. Blake. 1998. 395 pp.

37 **GIACALONE-RAMAT, Anna and Paul J. HOPPER (eds.):** The Limits of Grammaticalization. 1998. vi, 307 pp.

36 **NEWMAN, John (ed.):** The Linguistics of Giving. 1998. xv, 373 pp.

35 **GIVÓN, T. (ed.):** Grammatical Relations. A functionalist perspective. 1997. viii, 350 pp.

34 **GIVÓN, T. (ed.):** Conversation. Cognitive, communicative and social perspectives. 1997. viii, 302 pp.

33 **FOX, Barbara A. (ed.):** Studies in Anaphora. 1996. xii, 518 pp.

32 **BYBEE, Joan and Suzanne FLEISCHMAN (eds.):** Modality in Grammar and Discourse. 1995. viii, 575 pp.

31 **GERNSBACHER, Morton Ann and T. GIVÓN (eds.):** Coherence in Spontaneous Text. 1995. x, 267 pp.

30 **DOWNING, Pamela A. and Michael NOONAN (eds.):** Word Order in Discourse. 1995. x, 595 pp.

29 **KAHREL (PJK), Peter and René van den BERG (eds.):** Typological Studies in Negation. 1994. x, 385 pp.

28 **GIVÓN, T. (ed.):** Voice and Inversion. 1994. viii, 402 pp.

27 **FOX, Barbara A. and Paul J. HOPPER (eds.):** Voice: Form and Function. 1994. xiii, 377 pp.

26 **LORD, Carol:** Historical Change in Serial Verb Constructions. 1993. x, 273 pp.

25 **SVOROU, Soteria:** The Grammar of Space. 1994. xiv, 290 pp.

24 **PERKINS, Revere D.:** Deixis, Grammar, and Culture. 1992. x, 245 pp.

23 **KEMMER, Suzanne:** The Middle Voice. 1993. xii, 300 pp.

22 **PAYNE, Doris L. (ed.):** Pragmatics of Word Order Flexibility. 1992. viii, 320 pp.

21 **DOWNING, Pamela A., Susan D. LIMA and Michael NOONAN (eds.):** The Linguistics of Literacy. 1992. xx, 334 pp.

20 **CROFT, William, Suzanne KEMMER and Keith DENNING (eds.):** Studies in Typology and Diachrony. Papers presented to Joseph H. Greenberg on his 75th birthday. 1990. xxxiv, 243 pp.

19:2 **TRAUGOTT, Elizabeth Closs and Bernd HEINE (eds.):** Approaches to Grammaticalization. Volume II. Types of grammatical markers. 1991. xii, 558 pp.

19:1 **TRAUGOTT, Elizabeth Closs and Bernd HEINE (eds.):** Approaches to Grammaticalization. Volume I. Theoretical and methodological issues. 1991. xii, 360 pp.

18 **HAIMAN, John and Sandra A. THOMPSON (eds.):** Clause Combining in Grammar and Discourse. 1988. xiii, 428 pp.

17 **HAMMOND, Michael, Edith A. MORAVCSIK and Jessica WIRTH (eds.):** Studies in Syntactic Typology. 1988. xiv, 380 pp.

16 **SHIBATANI, Masayoshi (ed.):** Passive and Voice. 1988. xi, 706 pp.

15 **AUSTIN, Peter (ed.):** Complex Sentence Constructions in Australian Languages. 1988. vii, 289 pp.

14 **HINDS, John, Shoichi IWASAKI and Senko K. MAYNARD (eds.):** Perspectives on Topicalization. The case of Japanese WA. 1987. xi, 307 pp.

13 **Never published.**

12 **NEDJALKOV, Vladimir P. (ed.):** Typology of Resultative Constructions. Translated from the original Russian edition (1983). Translation edited by Bernard Comrie. 1988. xx, 573 pp.

11 **TOMLIN, Russell S.:** Coherence and Grounding in Discourse. Outcome of a Symposium, Eugene, Oregon, June 1984. 1987. viii, 512 pp.

10 **RANSOM, Evelyn N.:** Complementation: its Meaning and Forms. 1986. xii, 226 pp.

9 **BYBEE, Joan:** Morphology. A Study of the Relation between Meaning and Form. 1985. xii, 235 pp.

8 **SLOBIN, Dan I. and Karl ZIMMER (eds.):** Studies in Turkish Linguistics. 1986. vi, 294 pp.

7 **CRAIG, Colette G. (ed.):** Noun Classes and Categorization. Proceedings of a symposium on categorization and noun classification, Eugene, Oregon, October 1983. 1986. vii, 481 pp.

6 **HAIMAN, John (ed.):** Iconicity in Syntax. Proceedings of a symposium on iconicity in syntax, Stanford, June 24–26, 1983. 1985. vi, 402 pp.

5 **RUTHERFORD, William E. (ed.):** Language Universals and Second Language Acquisition. 1984. ix, 264 pp.

4 **CHISHOLM, William, Louis T. MILIC and John A.C. GREPPIN (eds.):** Interrogativity. A colloquium on the grammar, typology and pragmatics of questions in seven diverse languages, Cleveland, Ohio, October 5th 1981-May 3rd 1982. 1984. v, 302 pp.

3 **GIVÓN, T.:** Topic Continuity in Discourse. A quantitative cross-language study. 1983. vi, 492 pp.

2 **HAIMAN, John and Pamela MUNRO (eds.):** Switch Reference and Universal Grammar. Proceedings of a symposium on switch reference and universal grammar, Winnipeg, May 1981. 1983. xv, 337 pp.

1 **HOPPER, Paul J. (ed.):** Tense-Aspect. Between semantics & pragmatics. 1982. x, 350 pp.