Typological Studies in Language 83

Formulaic Language

VOLUME 2

Acquisition, loss, psychological reality, and functional explanations

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

John Benjamins Publishing Company

This book is the second of the two-volume collection of papers on formulaic language. The collection is among the first in the field. The authors of the papers in this volume represent a diverse group of international scholars in linguistics and psychology. The language data analyzed come from a variety of languages, including Arabic, Japanese, Polish, and Spanish, and include analyses of styles and genres within these languages. While the first volume focuses on the very definition of linguistic formulae and on their grammatical, semantic, stylistic, and historical aspects, the second volume explores how formulae are acquired and lost by speakers of a language, in what way they are psychologically real, and what their functions in discourse are. Since most of the papers are readily accessible to readers with only basic familiarity with linguistics, the book may be used in courses on discourse structure, pragmatics, semantics, language acquisition, and syntax, as well as being a resource in linguistic research.

ISBN 978 90 272 2996 0

TSL 83 Corrigan et al. Formulaic Language VOLUME 2

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 83

Formulaic Language. Volume 2. Acquisition, loss, psychological reality, and functional explanations.
Edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen M. Wheatley

Formulaic Language

VOLUME 2

Acquisition, loss, psychological reality, and functional explanations

Edited by

Roberta Corrigan
Edith A. Moravcsik
Hamid Ouali
Kathleen M. Wheatley

University of Wisconsin-Milwaukee

John Benjamins Publishing Company
Amsterdam / Philadelphia

Library of Congress Cataloging-in-Publication Data

Formulaic Language : volume 2 : acquisition, loss, psychological reality, and functional explanations / edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen M. Wheatley.
p. cm. (Typological Studies in Language, ISSN 0167-7373 ; v. 83)
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 2996 0 (Hb; alk. paper) – ISBN 978 90 272 2997 7 (set : alk. paper)

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

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.

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 esthetic 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

Acquisition and loss

Repetition and reuse in child language learning

Colin Bannard & Elena Lieven
Max Planck Institute for Evolutionary Anthropology

Abstract

Language is a social tool that we learn by observing and interacting with others. We can use language to communicate with others because of the way in which it is shared as a set of conventions across our communities via this process of exchange. In this paper we will show how the formulaicity of language follows directly from this very simple and intuitive view. Formulaicity can be thought of as language reuse. We will make the case that language reuse is not simply one aspect of linguistic communication but rather its very basis. We first consider evidence for the existence of formulas and their usefulness in processing. Next we report a study which analyses two-year-old children's novel utterances in terms of the combination of recurring strings of speech that they have either used or heard previously. Finally we discuss experimental results that suggest that children may indeed store frequent strings as chunks that can be retrieved as a whole.

1. What is a speech formula and why?

It seems to us that the best definition of the speech formula is a statistical one – it is a multiword piece of language that occurs a lot. One might want to slightly

modify this to say that it is a piece of language that occurs a lot relative to the rate of occurrence we would expect given its component words (see Manning & Schütze 1999, chapter 5 for a discussion of this point), and we accept the argument (made e.g., by Wray 2002) that due to variation in individual experience what functions as a formula for one individual might not have a high frequency across the speech community as a whole. Nevertheless it seems that if we are to get anywhere in discussing formulaicity then a necessary first step is to take a snapshot of language and look at how much and what kinds of repetition we can see. A somewhat coarse but still striking view is provided by figure 1. The snapshot we use here is the 89 million word written component of the British National Corpus (Burnard 2000). The figure plots the natural logarithm of the frequency of each string of words encountered against the natural logarithm of the position of each string in a ranked list of these substrings. It is a long acknowledged pattern that the relationship between rank and frequency follows something close to a power law distribution (e.g., Zipf 1949). What is interesting about this distribution is that it means that a few events in language recur frequently, while most are somewhat rare. This seems to be a universal law that can be observed in any sample of any language. The interesting thing in the context of this volume is that it applies not just for words but for multiword sequences too.

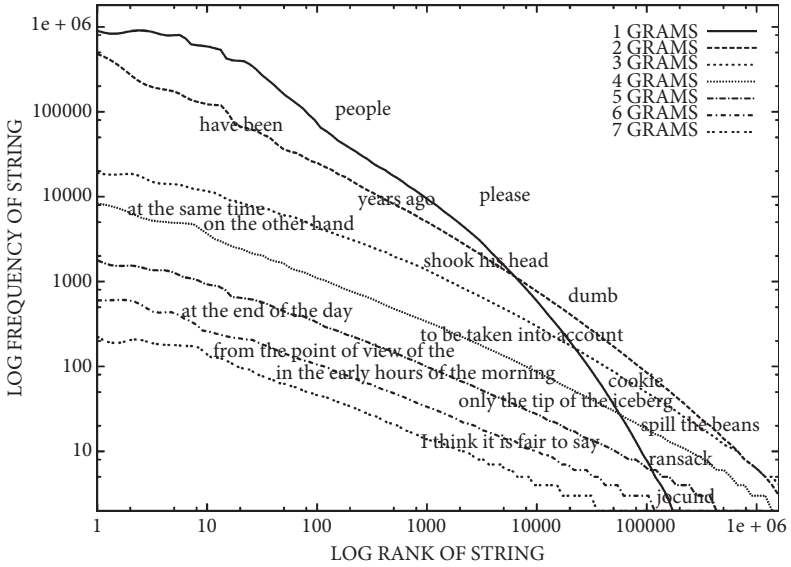


Figure 1. Repetition and reuse in the British National Corpus.

This Zipf plot, then, gives us a nice visual representation of formulaicity in language. A number of multiword sequences (e.g., "at the same time", "shook his head", "years ago") can be seen to occur with greater frequency than core parts of the single word vocabulary of English (e.g., "dumb", "cookie"). This seems to be evidence that these phrases have some privileged status. So what is this status exactly? There are a number of explanations for Zipf's law in the literature. Zipf's original claim was that it was the result of a compromise between the need of the speaker and the hearer to minimize their processing cost (what he called the "principle of least effort"). The speaker wants to minimize the diversity of what is produced but the hearer needs some diversity in order to disambiguate. The recurrent transmission of language over this channel results in a situation where there is enough repetition to make things easy for the speaker and enough diversity to avoid ambiguity. A formal model showing that this is indeed what results from such a process of transmission has recently been provided by Ferrer i Cancho & Sole (2003). For those who do not like Zipf's cognitive model, a purely formal proposal was made by Simon (1955).[1] In this account the likelihood of any word being repeated is exactly a function of how often it has been encountered before. If language is generated according to this principle with a certain amount of probability space being held back for novel words to avoid stagnation, then the resulting distribution is very close to that we observe in natural language. A shared assumption of both these approaches is that the more a word or phrase has been heard before the more it will be heard in the future. This kind of power law distribution is not peculiar to language and can in fact be observed in any number of social phenomena (e.g., links found on the internet, academic paper citations; see e.g., Barabasi 2002 for an accessible survey). The process that is thought to drive this is "preferential attachment" – people prefer to link to websites and cite papers that have been linked to or cited before. Similarly in language, they prefer to use words that have been used by others before. As can be seen from the above figure the same basic law seems to apply for multiword sequences. Given the way in which language functions as a conventional communication system this seems like a highly effective strategy by which to achieve successful communication. The outcome of this process is precisely the topic of this book – a community that generates speech according to some process of preferential attachment is a community in which speech will to some extent be repetitive and formulaic.

1. See Baayen (2001) for discussion of various models. For a different perspective see the claim of Miller (1957) that Zipf's law is actually just the result of the distribution of silence in language. Then see Howes (1968) for a criticism of this argument.

1.1 Repetition and analogy

The question that interests us as psychologists, then, is how and why this situation comes about. What is it that drives this move to formulaicity? We take a usage-based perspective on language. From this perspective reuse and repetition is central. While this approach has been reviewed often elsewhere (see e.g., Tomasello 2003), we will briefly review the main points here. The basic assumption of usage-based linguistics is that language is functional at all levels and that all parts of language are form-meaning mappings. The most basic form-meaning mapping is of course the mapping between a series of words and some entity or event. The speech formula is thus the simplest example of a multiword linguistic symbol. Crucially, however, the role of reuse doesn't end there. Much of language is productive. While some linguists in the generative tradition have been guilty of overstating how productive language is, there is no doubt that we are able to produce and understand utterances we have not produced or heard before. What distinguishes usage-based linguists from generativists is that rather than accounting for productivity in terms of some language specific, pre-experiential mechanism such as innate linking rules, we assume that our ability to understand novel utterances derives from the fact that they are like utterances we have heard before. Whether one assumes that the learner forms a set of abstract constructions that can be applied directly, or that language is always entirely exemplar-based, or that the truth is somewhere in the middle (see Abbot-Smith & Tomasello 2006 for discussion) the ultimate conclusion is the same – language understanding occurs through an analogy, either performed online or already implicit in categorical knowledge, between the input and one's prior linguistic experience. For this reason meaning in language is always derived to some extent from repetition – we can only understand new utterances to the extent that they are like what we have heard before. Accordingly, from a usage-based perspective the task of language learning can be thought of as learning to appropriately reuse the language that one hears.

1.2 Segmentation, data compression and efficiency through redundancy

Let's start right at the beginning and consider the obstacles faced by the infant at the outset of language learning. The job of the infant in development is to learn to understand and predict his/her environment. One prominent recurrent aspect of this environment are the vocal sounds and gestures that others make. To understand this s/he will need to discover structure and regularities. Both usage-based and generativist perspectives have to account for the way that the infant notices regularities in the speech stream such that s/he develops a vocabulary. The problem faced by the child of taking an undifferentiated stream of sounds and determining the words or units of meaning has attracted considerable attention. A number of these have

focused on word boundary cues available in the input, either phonetic (e.g., Lehiste 1971), or prosodically marked (e.g., Grosjean & Gee 1987). More recently there has been a very large body of evidence that children are able to perform segmentation by simply observing the regularity with which sounds co-occur in the language. The basic assumption of these models of segmentation is that children track the frequency with which particular sounds precede others. Tracking these regularities allows the child to segment the stream. Sounds that occur together frequently (or where the probability of seeing a particular sound given a preceding sound is high) are taken to be words or word-components. Conversely, word boundaries can be posited at points where low frequency or low probability transitions between sounds are observed. Evidence for this has come from a number of studies involving adults (Saffran, Newport & Aslin 1996), children (Thiessen & Saffran 2003), and infants as young as 8 months (Saffran, Aslin & Newport 1996).

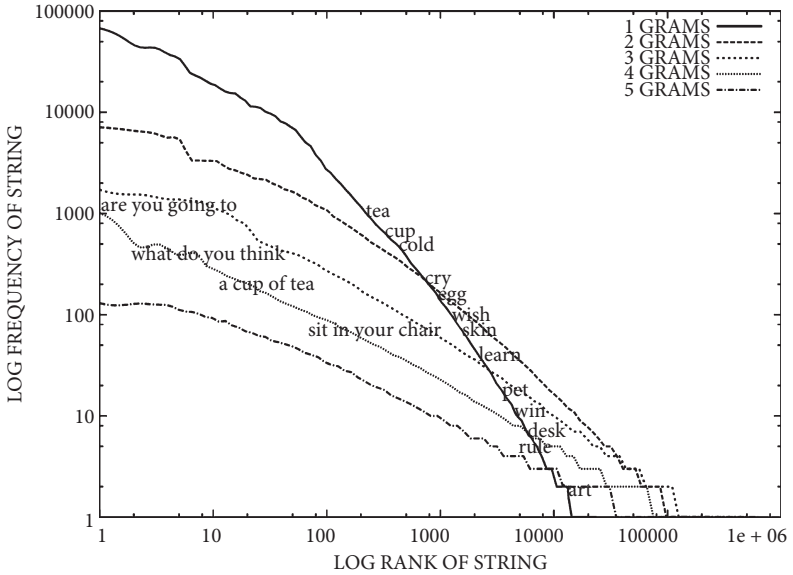


Figure 2. Repetition and reuse in child-directed speech.

If children do indeed perform probabilistic segmentation then the next question is what kind of sequences would they begin to pick out in their linguistic environments? In considering this it will be useful to look at another Zipf plot. Figure 2 presents a picture of the input encountered by the child. What is plotted here is

the frequency of strings of different lengths encountered in a substantial collection of mother's speech. The corpus used is a part the Max Planck Dense Database Corpus, collected by the Max Planck Child Study Centre, Manchester. This section contains the speech addressed to and produced by a single child, Brian, between the ages of two and five (1.7 million words of maternal input over 0.3 million utterances).[2] What is interesting about this plot is again the repetition. It is not only single words that occur with very high frequency but also longer strings. And again these occur with greater frequency than much of the single word vocabulary encountered. Take the word "pet" which occurs 12 times. There are strings as long as five words that occur with greater frequency than this. As a second example, take the word "tea", which occurs 1000 times. There are strings of two, three and even four words in length that occur with the same or greater frequency than this. Crucially there are also strings as long as four words ("a cup of tea") that fall only just below "tea" on the logarithmic scale. From the point of view of learning this means that a child will frequently encounter the same multiword sequences.

Thus if children are using statistical information to extract meaningful units from the speech stream we would expect that they would acquire multiword sequences as well as individual words. This conclusion has had substantial support from modelling work. There have been a large number of computational models of how segmentation using distributional information might work. Models have been built to segment artificial strings (e.g., Elman 1990), child-directed speech (Brent & Cartwright 1996), transcribed adult conversational speech (Cairns, Shillcock, Chater & Levy 1997), and written texts (Kit & Wilks 1999). A consistent finding of these models has been that while this approach to segmentation results in a large number of what are conventionally considered to be words, any approach that does not oversegment also seems to result in a large number of multiword sequences.

So given the way that we know children perform segmentation we would expect that a child might extract multiword sequences. What evidence is there that this is in fact what they do? In a seminal paper, Peters (1977) described the speech of a child she called Minh, whom she recorded for up to an hour a week from the age of 0;7 months to 2;3. She recounts how from the age of 1;5, he demonstrated a style of speech that she calls "mush-mouth". This was mumbled production, where

2. Brian lived in a large metropolitan area in England and came from a middle-class background. His mother made one-hour tapes of herself and Brian in natural interactions in their home. His father, other adults and a research assistant were also occasionally present for the recordings. Recordings span the three years between Brian's second and fifth birthdays. For the first year, one month and fourteen days, recordings were made for 5 days each week. For the remainder of the time, recordings were made five days per week for one month per year. These recordings were then transcribed in CHAT format (MacWhinney 2000).

phrases were approximated by their intonational contours, seemingly aiming at the reproduction of whole sentences or utterances rather than individual words. Many of his utterances were very frequently repeated phrases for which context was very useful in determining meaning, such as *Look at this!*, *What's that?* or *Open the door*. However he also attempted to reproduce set repetitive utterances, using an established filler sound for the unclear parts. Peters suggests that this kind of "gestalt" learning of speech is an important part of language which tends to be overlooked by researchers who look in child speech for the same units and levels that are supposed for adult language.

It seems that at early stages of development, children use multiword sequences in language production. The question that we might want to ask then is whether this is purely a developmental stage. It is clear that children do eventually acquire the ability to use individual words. Might we not expect that the multiword sequences would begin to erode at that point? This would be consistent with the traditional assumption in generative linguistics that the basic units of language are individual words and morphemes. The assumption here seems to be that only storing words is the optimal strategy if one wants to produce and comprehend language. However, we think that there is a strong argument that this is not the case. A recurrent theme in the generative tradition which has appeared in numerous guises is the idea that the job of the linguist is to provide the simplest possible way of accounting for the data. As Goldsmith (in press) points out, this seems to be similar in some respects to the goal of minimum description length modelling in statistics (Rissanen 1989) and accordingly to data compression in computer science (Salomon 2000). The task in data compression is to provide the most efficient way of encoding some data (either for storage on one's hard drive or transmission over the internet). This means finding the description of the data that takes up the smallest amount of space. For ease of explication the kind of data that we are considering here will be text. If one thinks of a text as being a string S then one can think of text compression as the problem of finding the string S^1 that is a shorter (or ultimately the shortest) string from which one can reconstruct S. This can be done by exploiting redundancy in the text. Say S is the text "to be or not to be or to be or not to be". This consists of 40 characters. However, the string contains many repeated sequences, such as the words "to" and "be" and the multiwords "to be" and "not to be". One can exploit this redundancy by assigning a unique ID to these substrings and then replacing the substring with this ID as a placeholder. When one has to view the text, the substring can be substituted in for the ID. The complication in doing this is of course that one will need to add additional information to the string in the form of the dictionary that records which ID matches to which substring. However if there is sufficient redundancy (repetition) in the text then creating the dictionary will still be efficient. Say one wanted

to exploit the repetition of "to be" and of "not to be" then the dictionary might be of the form "1=to be;2=not to be", and one could then create the text S^1 as follows "1 or 2 or 1 or 2#1=to be;2=not to be". This consists of a total of 36 characters. One can generate text S from text S^1 by a process of simple dictionary look up and substitution. By choosing this encoding one would use 10% less disk space than if one were to store S directly.

It may seem a somewhat unlikely comparison but text compression is analogous to the problem of language learning in some crucial respects (see Chater & Vitanyi 2003 for a proposal that simplicity is a general principle of human cognition). Speakers of a language need to be able to comprehend the utterances that they encounter and generate their own in such a fashion that they can in turn be understood. By simply storing the individual words a learner would be able to do this. However, if they are to thrive in the world they need to optimize their gains. The optimal learning strategy would be to learn a representation that allows them to account for the data most efficiently, which is to say the representation that allows them to understand and/or produce language with the fewest processing steps, without increasing the size of their mental representation unnecessarily. While the "to be or not to be" example we gave above was somewhat artificial, the basic idea of exploiting redundancy and creating a dictionary in order to compress text is the basis for many text compression algorithms used in software that we use on a daily basis (e.g., Ziv & Lempel 1978). Research in this area has consistently shown that the most efficient way to encode data is not just to include multiword sequences in the dictionary, but to employ what Ziv & Lempel (1965) call "variable length coding". So in our example we included "to be" and "not to be" in the dictionary. This might appear to contain redundancy since the first entry can already handle part of the second entry. However, the statistics of natural language text mean that if one wants to optimally account for the text data one needs to have both entries in the dictionary. Many years of research into the statistics of natural language in computer science have shown that including what from a generativist perspective might look like redundancy is in fact the most efficient way to proceed.

We have seen that there are good theoretical reasons for thinking that children will continue to store multiword chunks as language learning progresses. We now turn to the evidence for this in children's language learning and ask how children might use these chunks to develop their language.

1.3 Children learn chunks from what they hear

When children hear a string, it may sometimes remain as an unanalyzed chunk mapped as a whole to a particular function. An example might be *What's that?* which is highly frequent in speech directed to English 2-year-olds. The child is very unlikely to have analyzed this in such a way that the '*s* is related to copula BE,

nor is s/he likely to have a meaning for the utterance which includes the idea that the interlocutor really does not know what the object referred to is. Rather s/he will map the whole chunk to some meaning more like "Display to me your knowledge of the name of that object". An adult, on the other hand while still retaining the whole string as a formula that can be produced (see, for instance, Bybee & Scheibman 1999) would probably also 'know' the relationship between this string and others such as *What IS that?* and *I know what that is*. An early example of how such unanalyzed chunks might be used by children is provided by Clark's (1974) study of her son's early utterances (example 1):

Example 1: from Clark (1974)

MOT:	We're all very mucky
CHI:	I all very mucky too
MOT:	(preventing child from putting arms in wrong end of sleeve): That's upside down
CHI:	No, I want to upside down

Note that in this example the child is productively using a chunk from the preceding discourse together with other elements from his own linguistic system. This raises the second important process associated with the learning of chunks – that they may be broken down in a process that develops increasingly general 'slots'.

2. Chunks may become analyzed

Researchers have drawn attention to two processes through which this can occur. First, as documented extensively by Peters (1997), typological features of the language being learned may lead to analysis: for instance stress patterns may identify components as separate, leading children to create 'placeholders' for particular morphemes (e.g., Aksu-Koc & Slobin 1985, postverbal morphemes in Turkish; Demuth 1992, for noun class prefixes in Sesotho) and to try to identify meanings for them in a process Tomasello calls 'blame assignment' (Tomasello 2003). Secondly, if children hear many strings that have overlapping phonological and positional consistency together with type variation, they may form a frame and slot construction. An example is the very frequent *Where's X?* string used by English-speaking parents to their children and often learned very early by the children. Here, again, the frame part of the construction is almost certainly not analyzed into a wh-question word (linked to others) and a form of the copula, but the slot is open. Different children may have narrower or broader categories for what can fill the slot – perhaps only a group of people for one child but a wider range of 'objects' for another. And this leads to the third process involved in the formation of slots – their development into more general and schematic categories.

2.1 Slots provide the basis for developing more general categories

In his important 1976 monograph, Braine analyzed a number of corpora of early child speech and showed that as well as strong similarities in what children talked about across languages, they were also similar in using what he called positional patterns in which one word reoccurred in the same position while the other word in the utterance came from a much more variable set, for instance *More + X, Want + X* (Braine 1976). Braine's main interest was in analyzing the underlying semantic relations expressed by these patterns but his work was seminal in leading researchers to see that variation in a slot might be the basis for a developing category. A more recent example comes from Tomasello's monograph (Tomasello 1992) on his daughter's early utterances with verbs. He argued that at this early age, his daughter had a number of low-scope constructions based on individual verbs but that argument roles were not generalized beyond the individual verb. Thus if the child used the verb *hit* in a construction with two arguments, she would know where to place the 'hitter' and 'hittee' but this knowledge did not extend more generally to subject and object argument roles or even to more semantic categories like 'doer' and 'done to'. Tomasello argued that these syntactic roles develop slowly and that children may only develop a fully schematic transitive construction some time in the second half of their third year.

Thus children are seen as building up a network of constructions in which relations between constructions and parts of constructions are continuously developing both in their levels of entrenchment and their connectedness with other constructions. An example is provided by Pine, Lieven & Rowland's (1998) analysis of children's early utterances with verbs in which they suggested that, as well as 'verb island' constructions based around particular verbs, children also developed constructions that had slots for verbs. Examples are *I can't X* and *I'm X-ing.* This does not necessarily mean that the child has a fully schematic verb category – as Clark (1996) argued, children may have sub-categories of verbs, for instance, punctual versus durative verbs. But it does illustrate the way in which children are seen as building up an inventory of constructions. As these become analyzed into parts, they also become connected to other constructions and parts of constructions in an increasingly complex network, which allows for the development of more schematic constructions and categories.

2.2 Productivity and creativity

As we pointed out earlier, human language provides the basis for the production of novel utterances by speakers – utterances they have never heard or produced before – which may map to meanings that are also more or less novel. In this sense of creativity, children are creative with language from the moment that they use a

label in a context they have not heard it before, for instance to refer to a new, previously unseen teddy as *teddy* or, in Braine's (1976) example to say *more wet*, when asking for the bath taps to be turned on again. The issue is not, then, whether children are able to produce novel utterances but the basis on which these utterances are generated (see, for instance, Cameron-Faulkner, Lieven & Theakston 2007 on one child's early patterns of negation). What is the initial level of abstraction from which children construct their utterances? From the perspective of generative grammar, while performance factors may affect the process of production, there is a fundamental sense in which an underlying grammar based on abstract and universal linguistic representations is involved (O'Grady 1997). From a usage-based perspective, the schematicity and abstraction of adult grammar arises through the developmental process of building an inventory of constructions. This process is centrally influenced by patterns of meaning and frequency in what children hear and produce (Tomasello 2003).

One major problem with attempting to identify creative utterances and the basis of their productivity is the difficulty of determining whether any particular utterance has been rote-learned or generated from smaller parts. This is particularly problematic because the sampling frames in most child language studies are extremely limited. Thus corpora typically consist of one to two hours of speech recorded every 2–3 weeks. Calculated on the basis that the child is awake for about 10 hours a day, this constitutes something like 1–2% of what s/he hears and says. This makes for major problems when attempting to measure degrees of entrenchment, productive schemas and novel utterances (Rowland, Fletcher & Freudenthal, 2008; Tomasello & Stahl 2004). Recently we have been collecting much denser corpora of between 5–10 hours per week over more or less extended periods which, we calculate, gives a coverage of between 7–20% of the child's waking hours (see Cameron-Faulkner et al. 2007; Lieven, Behrens, Speares & Tomasello 2003; for studies using the first two of these 'dense databases'). In a method we call 'traceback' we have attempt to identify novel utterances and to find their possible basis in what the child has said or heard before (for more details of this method and its origins, see Dąbrowska & Lieven 2005; Lieven et al. 2003).

2.3 The 'traceback' method

Using this method we first divide each child's corpus into a main and a test corpus. The idea is to trace novel utterances in the test corpus back to strings in the main corpus from which they could have been constructed. In the present study, six weeks of recordings were used with the last two hours acting as the test corpus and the preceding recordings acting as the main corpus (either produced by the child or an adult). In the test corpus we identify all the child's novel utterance

types ('target utterances') and then look for strings in the main corpus that have shared lexical material – these we called 'component units'. To be identified as a component unit the string has to occur at least twice (excluding immediate imitations and repetitions) and two types are defined: fixed phrases and schemas with slots. A fixed phrase is any continuous string of words corresponding to a 'chunk' of semantic structure (i.e., designating a REFERENT, PROCESS, LOCATION, DIRECTION, ATTRIBUTE, etc.) which occurs at least twice in the main corpus. The string does not have to occur in isolation – so the following two utterances are regarded as evidence that the expression *make a cake* is available to the child as a unit (fixed phrase) which can be placed into a PROCESS slot:

Example 2: Fixed phrase as component unit (Eleanor 2;0)
CHI: oh lets make a cake.
CHI: Mama you make a cake.

If a string occurred that matched the novel utterance in the same way, with variation in the same position, this was identified as a schema with a slot. A slot was established if at least two different expressions belonging to the same broad semantic category occurred in the same position in the schema (see table 1). Thus, the following two utterances are evidence for the schema *I got no REFERENT on* which is available to the child:

Example 3: Schema with slot as a component unit (Annie 2;0)
CHI: I got no sock-s on.
CHI: I got no trouser-s on.

Table 1. Types of semantic slots

Type of slot	Example	Utterances	Schema with slot
REFERENT	CHI	more choc+choc **on there**.	**REFERENT on there**
	CHI	Bow-'s food **on there**.	
PROCESS	CHI	**I want to** get it.	**I want to PROCESS**
	MOT	and **I want to** talk to you about the park.	
ATTRIBUTE	CHI	Pilchard there **he's** hungry @sc toast.	**he's ATTRIBUTE**
	CHI	**he's** upside+down.	
LOCATION	CHI	**I sit** on my Mummy-'s bike.	**I sit LOCATION**
	CHI	**I sit** there.	
DIRECTION	CHI	**going** under bridge.	**going DIRECTION**
	CHI	**going** down.	
POSSESSOR	INV	this is my **favourite**.	**POSSESSOR favourite**
	MOT	yeah it's your **favourite** that one, isn't it?	
UTTERANCE	CHI	there's sand **on it**.	**UTTERANCE on it**
	INV	a big flower **on it**.	

For each target utterance, all potential component units are identified in the main corpus and an attempt is then made to derive the target utterance using two operations defined as SUBSTITUTE and ADD. SUBSTITUTE allows the placement of a component unit into the slot of a schema and ADD allows the placement of component units to one or other end of an utterance.[3]

Example 4: (Fraser, 2;0)
Target utterance: A big mess.
Components: 1. a big REF
2. big mess

The fixed phrase *big mess* is substituted into the REFERENT slot in the schema ***a big REF*** in order to derive the target utterance *a big mess* using only one operation. Note that, in the case of this example, there is shared material between the schema and the fixed string. This does not have to be the case provided the fixed string is compatible with the semantics of the slot. The following is an example of a multi-operation derivation:

Example 5: (Eleanor 2;0)
Target utterance: Don't touch those flakes
Components: 1. don't touch REF
2. those REF
3. flakes

Here *those REF* is substituted into *Don't touch REF* and, in a second operation, *flakes* is substituted in the REF slot in *Don't touch those REF.*

In deciding between component units, the following rules were observed:

1. The longest possible schemas were used.
2. The slots were filled by the longest available units.
3. The minimum number of operations were taken.
4. In judging whether a string was a potential candidate as a slot filler, the semantics of the slot and of the filler string has to match.

Component units were identified using a computer program but all results established by the program were manually checked and revised where necessary, taking semantics into account. Semantic coding was done, after extensive training, by two research assistants. For reliability, 20% of all tracebacks were coded twice. Agreement was high (kappa = 0.89)

3. ADD was only allowed if the component unit could, in principle, go at either end of the utterance. This turned out to be largely confined to vocatives and was extremely rare, making up only 1–2% of all the operations – we do not consider it further here.

Using this method, we examine the corpora of 4 children (Brian,[4] Fraser, Annie and Eleanor) collected from their second birthdays to six weeks later on a schedule of one hour of recording on 5 days per week. The children came from a large city in the north of England. The families were from middle-class backgrounds. Three were only children at the time of recording (Brian, Annie and Eleanor) while Fraser had an older brother. The mothers were trained to make four of the five weekly recordings while a research assistant visited to make the fifth recording on video. The most typical activities were playing or having a snack. Research assistants transcribed all of the tapes in SONIC CHAT format (MacWhinney 2000). All speech was transcribed with the exception of the speech not directed to the child (i.e., speech between adults, telephone calls etc.).

The children's data were coded for imitations, self-repetitions, partially intelligible and incomplete utterances and routines. The mother's data were coded for the last three mentioned items. Utterances were coded as imitations and self-repetitions if they were exact or reduced repetitions (exact repetitions of a subpart of the utterance) of one of the last 5 utterances of the mother or child. Each utterance was linked to the appropriate sound file which makes it possible to listen to the utterance if required. The children's utterance types in the last two hours of recording were the target utterances of the test corpus and the children's and mothers' utterances in remaining recordings were used as the main corpus.

The children's mean length of utterance (MLU) was measured across all their utterances in the corpus. At 2;0, and for children learning English, this is a reasonable method of broadly assessing relative levels of language development. In increasing order, the children MLUs were: Brian, 1.65, Fraser 1.8, Annie, 2.19 and Eleanor, 2.22.

2.3.1 *Results using the traceback method*

Utterance construction

Table 2 presents the results of the traceback showing the number of target utterances that were exact repetitions of entire strings said at least twice in the main corpus ('zero operations'), those that required only one operation of substitution or addition to a string in the main corpus to produce the target, those requiring more than one operation and those that could not be derived from the main corpus ('fails'). Fails occurred either when a particular word or string in the target utterance could not be found in the main corpus ('lexical fails') or when there was no matching schema with a semantically compatible slot ('syntactic fails').

4. This is the same child whose CDS corpus was described earlier.

Table 2. Number of operations required to derive target utterances in the test corpus from strings in the main corpus

%	Brian	Fraser	Annie	Eleanor
zero operations	41.1	40.5	24.7	24.9
single operations	40.2	43.8	48	35.9
multi operations	1.8	11.6	17.5	23.9
lexical fails	2.7	1.3	3.6	6.7
syntactic fails	14.3	3	6.1	8.6
Total number of target utterances	112	304	279	209

It can be seen that around 25 – 40% of the children's target utterances are exact repetitions of strings already produced in the main corpus and a further 36 – 48% could be derived by just one operation. The number of lexical fails is low (and if the child has said the word or string in the target utterance, s/he must, at some point, have heard it) and the number of syntactic fails is between 3 – 14%. Note also that, with increasing MLU, the number of exact repetitions goes down and the number of multi-operation derivations goes up. This makes sense if, as children develop their language, they depend less on simply repeating rote-learned strings and are able to command more sophisticated 'assembly operations'.

The vast majority of these operations consist of substitutions into REFERENT slots (60 – 90%, see Table 3). This reduces somewhat with increasing MLU but does not go lower than 60%.

Table 3. Percentages of different types of slots identified in schemas

%	Brian	Fraser	Annie	Eleanor
Referent	89,6	69,3	60,2	63,2
Process	2,1	9,8	15,4	15,8
Attribute	0	2	9,8	7,9
Direction	0	1,5	1,2	0,5
Location	0	1	4,9	2,1
Possessive	0	0,5	1,2	0
Utterance	8,3	16,1	7,3	10,5
Total no. of slots	48	205	246	190

When PROCESS slots were not filled with single verbs (between 27–50%), they were either filled with strings containing negation (e.g., *can't, don't, not*) or a direct object and/or a preposition. The percentage of PROCESS slots increases with

increasing MLU as do ATTRIBUTE and LOCATION slots but since the numbers of these slots at 2;0 are rather low both absolutely and relative to REFERENT slots, we shall examine the latter in somewhat more detail.[5]

The development of schematization in REFERENT slots

With increasing MLU, the children placed more complex material into the REFERENT slots (see table 4). Thus Brian, with the lowest MLU, filled 80% of his REFERENT slots with single nouns, while Fraser, the child with the next highest MLU filled only 60% with single nouns. However if we compare his REFERENT slots with those of the two girls, we can see that he shows less variety, using almost exclusively either the definite or indefinite determiner, while Annie and Eleanor show a greater range of different types of noun phrase.

Table 4. Percentages of different types of fillers for REFERENT slots

%	Brian	Fraser	Annie	Eleanor
single nouns	81	58.9	73.5	59
a/the + noun	2.7	31.9	12.9	16
other determiner + noun	0	2.8	10.6	12
a/the + adjective + noun	0	1.4	0	4
adjective + noun	10.8	0	0.8	4
noun's (possessive) noun	5.4	2.1	1.5	0
pronoun	0	2.8	0.8	5
Total no. of REF fillers	37	141	132	100

The relatively high proportion of REFERENT slots and fillers shown by all the children supports the suggestion that English-speaking children may develop a relatively abstract 'noun phrase' earlier than other grammatical categories. Evidence for this comes from experiments in which children were quite willing to use a new object label in constructions that they already had (e.g., *Wug gone* or *More wug*) but did not do this for novel verbs that they were taught (Tomasello, Akhtar, Dodson & Rekau 1997). If the form-function mapping between referents and their labels is already available to children before they start producing multiword utterances, then schemas with referent slots may well be the earliest to develop and this, we suggest, is what we are seeing here. Once such a slot is available the child can start to 'notice' and use more complex constructions

5. We are conducting an ongoing traceback of the children's utterances at 3;0, where the proportion of PROCESS slots is much higher.

within it – thus we see that the difference in overall language development as measured by MLU is associated with more complex types of noun phrases.

Further evidence for this idea comes from an analysis of the children's ability to ground their REFERENT slots (Langacker, 1987). Many of the schemas we derived were already 'grounded' in the sense that determiners were already present in the schema, e.g., *I want a REF, There's the REF, Give me the other REF* but many were not, e.g., *I want REF, There's REF, Give me REF*. We examined those schemas containing ungrounded slots to see whether the children, in filling the slots also grounded them (see Table 5). With increasing levels of language development, children were more likely to correctly ground ungrounded slots in schemas.

Table 5. Percentage of ungrounded REFERENT slots filled with ungrounded and grounded fillers

	% Ungrounded	% Grounded-Proper Names	% Grounded-other	Total no. of slots
brian	50	23	27	30
fraser	22	28	48	58
annie	22	30	48	64
eleanor	13	30	58	71

We would suggest that what we see here is a process by which, having first learned a number of exemplars, the child develops a form-function mapping between an identified slot in a schema and a category of referents. Having done so, s/he is in a position to identify more complex referent labels in the input and to develop the ability to use these.

There are many issues related to this that we cannot take further here but which are the subject of ongoing research. A number of cautions need to be raised. First is that we are comparing 4 different children at the same age to try to get an idea of development. Clearly we need to follow particular children as they develop to see whether the pattern identified here is indeed a developmental one for individual children (Lieven, Salamo & Tomasello, submitted). Secondly we should note that the numbers of utterances are not controlled for in the present study and this will in turn affect the identification of prior strings. Third, since we are tracing the target utterances back to the main corpus, we are constrained by what the child says in the last two hours of recording. In a more recent project, we are extracting a grammar from the utterances in the main corpus and seeing how well they can generate the target utterances in what might be called a 'trace forward' (Bannard, Lieven & Tomasello, in press).

3. Experimental evidence for multiword storage

Up to this point we have provided arguments on the basis of rational analysis and naturalistic data. We now want to turn to some new experimental data that supports the idea that children store and utilize multiword sequences beyond the early stages of development. Bannard and Matthews (2008) sought to test the assumption that children store utterances as wholes by testing memory for familiar sequences of words. Using the same data as can be seen in figure 2 we identified frequently-occurring chunks in the input (e.g., *sit in your chair*) and matched, infrequent sequences (e.g., *sit in your truck*). The items were controlled so that all component individual words and bigrams had matched frequencies.We tested preschoolers' ability to produce these sequences in a sentence repetition test. It was found that both 2 and 3-year-olds were significantly more likely to correctly repeat frequent than infrequent sequences. Moreover, 3-year-olds were significantly faster to repeat the first 3 words of an item if they formed part of a chunk (e.g., quicker to say *sit in your* when followed by *chair* than *truck*). It appears, then, that children do indeed have dedicated representations for sequences of more than one or two words. Interestingly, while the overall performance of the 3 year olds was better than that of the 2 year olds no significant interaction between age and frequency was found meaning there is no clear sign that the two year olds are relying on multiword representations any more than the three year olds. From this perspective it is interesting that there is also evidence for multiword storage into adulthood (e.g., Bannard & Ramscar 2007).

In the final section of this paper, we deal briefly with two possible objections to the approach we have outlined here, first whether this type of chunk learning can account for errors and second, whether it can be applied to the learning of languages that are typologically different to English.

4. Learning chunks and making errors

It might be thought that if children are segmenting strings from the input, they should not make errors. Children's errors are of great interest in the study of language acquisition for two reasons. First, errors of commission give insight into the underlying representations from which children are generating these errors. Second, children are clearly not just repeating exactly what they hear. However, as we have already seen in the discussion of grounding above, children might substitute a semantically compatible but grammatically incorrect item into a slot-and-frame pattern. Thus the learning of chunks can both protect

from error and create errors. Rowland (2007) shows, for instance, that children's wh-inversion errors are related to the frequency with which they hear particular wh-auxiliary strings in the input. Highly frequent strings are produced without error but when the child tries to produce a wh-question for which there is much less evidence, non-inversion errors are significantly more likely to occur. A second example comes from Freudenthal, Pine, Aguado-Orea & Gobet's (2007) modeling of the optional infinitive error in English, Dutch, German and Spanish. Optional infinitive errors occur when children use a non-finite form for the main verb rather than correctly marking for finiteness. Error rates are very different in these four languages with Spanish having very low rates of error and Dutch the highest. The authors used a learning model that learned from right to left and successfully managed to model the different error rates in the four languages. The reason for this is not hard to find: when Dutch uses complex finite verbs, the modal or auxiliary is in verb-second position while the non-finite form of the main verb goes to the end of the utterance. Hence a right to left learner will pick up large numbers of non-finite verbs before learning the finite forms. Given the recency bias known to affect all sorts of learning, this may well account for the prevalence of non-finite verbs in the utterances of children learning Dutch and German.

5. Typological differences

English is a very peculiar language with more rigid word order and less inflectional morphology than most of the world's languages. It might be thought, therefore, that the approach taken here would not be appropriate for these other types of languages. The work has not yet been done to answer this question, so we can only give a few pointers to the directions in which we would like to go on this. In terms of languages with a wider range of syntactic word order patterns, it is important to note, first, that what is grammatically possible and what speakers actually do may not be the same. A study of Russian and German child directed speech (Stoll, Abbot-Smith & Lieven, in press) found very high levels of repetition in the first 2–3 words of the utterances addressed to children by their mothers, though not as high as those in the English sample that they also analyzed or in that of Cameron-Faulkner, Lieven & Tomasello (2003). We do not yet know how this affects children's learning of these languages but the reasons for the high levels of repetition are not hard to find: they derive from the nature of interactions with young children: many questions (with limited numbers of question words); many demonstratives, much elicitation of display language, many imperatives. Before concluding that, because a language is not as rigidly word ordered as English,

children will not be hearing chunks and using them to segment utterances, we need to examine what children actually hear. A usage-based approach would seek to find form-function mappings for variations in word order and to do this with actual speech corpora.

Secondly there is the question of inflectional morphology. Just as children often sound quite sophisticated as speakers through learning chunks of language and a mapping to meaning so, we would suggest, children may often sound more morphologically sophisticated than would be the case if one was able to truly probe the productivity of their system. In principle just as chunks are large words, so a chunk can consist of a word plus (a) morpheme(s), only later to be segmented and slowly connected into a network of other inflections. For instance, Aguado-Orea (2004) showed in an analysis of the verbal inflections of two Spanish speaking children, that even when he controlled the particular verbs and inflections used, the adults showed more productivity with these inflections than did the children. He also showed that the error rate for inflectional marking varied widely with person, number and the frequency of the verb being marked: highly frequent verbs (for instance *quiero* = I want, accounting for over 50% of all first person uses) being significantly more likely to occur without error.

6. Conclusion

In this paper we have presented rational and empirical arguments for a view of language in which formulaicity is the result of simple mechanisms in learning and development. At the outset we showed the distribution of words and phrases that we see in a sample of natural language. We pointed out the rate of repetition and explained how this follows a long-acknowledged universal law that can be explained using very simple models of cultural transmission. We then explained how this transmission can be seen in the way in which language is learned. We provided a rational argument that the reuse by children of whole multiword sequences taken directly from the input represents a very efficient learning strategy. We then looked at how the nature of this reuse might change over development as children acquire more complex grammatical knowledge, and described a series of studies that show how children's grammatical development can in fact be accounted for in these terms. We presented some new experimental data that supports the idea that children use multiword formulas throughout development. We finally pointed to some of the challenges that still face our model and indicated some initial ways in ways in which this model of learning as reuse is beginning to be successfully applied to even the thorniest issues in language acquisition.

Acknowledgements

We would like to thank Michael Tomasello for comments on an earlier draft; Danielle Matthews for collaboration on the experimental work, and Anna Roby, Manuel Schrepfer, Elizabeth Wills and Roger Mundry for help in data collection and statistical analysis for the same; Ewa Dąbrowka who was central to developing the Traceback method; Dorothé Salomo and Katja Hummel, for many hours of careful work on the Traceback analysis; Kristin Wolter, Silke Harbodt, and Daniel Stahl for further analysis and statistical assistance; the Dense Database team in Manchester, Jeannine Goh, Ellie O'Malley and Dimitra Doumpioti; and finally, and most importantly, the children and families who contributed so much of their time to the collection of the 'Dense databases'.

References

Abbot-Smith, Kirsten & Michael Tomasello. 2006. Exemplar versus prototype models in syntactic acquisition. In Special issue of *The Linguistic Review: Exemplar-Based Models in Linguistics,* S. Gahl & A. Yu (Eds), 23: 275-290.

Aguado-Orea, Javier 2004. The acquisition of morpho-syntax in Spanish: Implications for current theories of development. Ph.D. dissertation, University of Nottingham.

Aksu-Koc, Ayhan & Dan I. Slobin. 1985. The acquisition of Turkish. In *The crosslinguistic study of language acquisition,* Vol. 1: *The data*, D.I. Slobin (Ed.), 839–878. Hillsdale NJ.: Lawrence Erlbaum Associates.

Baayen, Harald. 2001. *Word frequency distributions.* Dordrecht: Kluwer.

Bannard, Colin, Elena Lieven and Micahel Tomasello, submitted. Modeling children's early grammatical knowledge.

Bannard, Colin & Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science* 19, 241–248.

Bannard, Colin & Michael Ramscar. 2007. Reading time evidence for storage of frequent multiword sequences, Abstract in Proceedings of *Architectures and Mechanism of Language Processing Conference (AMLAP-2007)*, Turku, Finland

Barabasi, Albert. 2002. *Linked: The new science of networks.* Cambridge MA: Perseus.

Burnard, Lou. 2000. *User reference guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Braine, Martin D. 1976. Children's first word combinations. *Monographs of the Society for Research in Child Development* 41(1): 104.

Brent, Michael R. & Timothy A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61(1–2): 93–125.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics* 37(4): 575–596.

Cairns, Paul, Richard C. Shillcock, Nick Chater & Joe Levy. 1997. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology* 33: 111153.

Cameron-Faulkner, Thea, Elena Lieven & Anna Theakston, 2007. What part of no do children not understand? A usage-based account of multiword negation. *Journal of Child Language,* 34: 251–282.

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

Chater, Nick & Paul Vitanyi. 2003. Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Science* 7(1): 19–22.

Clark, Eve V. 1996. Early verbs, event-types, and inflections. *Children's language* Vol. 9. Carolyn E. Johnson & John H.V. Gilbert (Eds), 61–73. Mahwah NJ: Lawrence Erlbaum Associates.

Clark, Ruth. 1974. Performing without competence. *Journal of Child Language* 1(1): 1–10.

Dąbrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16(3): 437–474.

Demuth, Katherine. 1992. The acquisition of Sesotho. In *The crosslinguistic study of language acquisition*, Vol. 3, D.I. Slobin (Ed.), 557–638. Hillsdale NJ: Lawrence Erlbaum Associates.

Elman, Jeff L. 1990. Finding structure in time. *Cognitive Science* 14: 179–211.

Ferrer i Cancho, Ramon & Ricard V. Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences* 100: 788–791.

Freudenthal, Daniel, Julian Pine, Javier Aguado-Orea & Fernand Gobet. 2007. Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science* 31: 311–341.

Goldsmith, John. In press. Morphological analogy: Only a beginning. In *Analogy in grammar: Form and acquisition*, J.P. Blevins & J. Blevins (Eds), Oxford: OUP.

Grosjean, Francois & James P.Gee 1987. Prosodic structure and spoken word recognition. *Cognition* 25(1–2): 135–155.

Howes, David. 1968. Zipf's Law and Miller's Random-Monkey Model. *The American Journal of Psychology* 81(2): 269–272.

Kit, Chunyu & Yorick Wilks. 1999. Unsupervised learning of word boundary with description length gain. In *Proceedings of 3rd conference on computational natural language learning.*

Langacker, Ronald. 1987. Nouns and verbs. *Language* 63(1): 53–94.

Lehiste, Isle. 1971. The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* 51(6(2): 2018 – 2024.

Lieven, Elena, Heike Behrens, Jennifer Speares & Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30(2): 333–370.

Lieven, Elena, Dorothè Salamo & Michael Tomasello, in press. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics. Special Issue on Language Development.*

MacWhinney, Brian 2000. *The CHILDES project: tools for analyzing talk*. Mahwah NJ: Lawrence Erlbaum Associates.

Manning, Christopher & Hinrich Schütze.1999. *Foundations of statistical natural language processing*. Cambridge MA: The MIT Press

Miller, George A. 1957. Some effects of intermittent silence. *The American Journal of Psychology* 70(2): 311–314

O'Grady, William. 1997. *Syntactic development*. Chicago IL: The University of Chicago Press.

Peters, Ann. 1977. Language learning strategies: Does the whole equal the sum of its parts? *Language* 53(3): 560–573.

Peters, Ann. 1997. Language typology, prosody and the acquisition of grammatical morphemes. In *The crosslinguistic study of language acquisition,* Vol. 5: *Expanding the contexts*, D.I. Slobin (Ed.), 135–197. Mahwah NJ: Lawrence Erlbaum Associates.

Pine, Julian M., Elena Lieven & Caroline F. Rowland. 1998. Comparing different models of the development of the English verb category. *Linguistics* 36(4): 807–830.

Rissanen, Jorma. 1989. *Stochastic complexity in statistical inquiry.* Singapore: World Scientific.

Rowland, Caroline. 2007. Explaining errors in children's questions. *Cognition* 104, 106–134.

Rowland, Caroline, Sarab Fletcher & Daniel Freudenthal. 2008. How big is big enough? Assessing the reliability of data from naturalistic samples. In *Trends in Corpus Research: Finding structure in data,* Heike Behrens (Ed.), Amsterdam: John Benjamins, 1–24.

Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294): 1926–1928.

Saffran, Jenny R, Elissa L. Newport & Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35: 606–621.

Salomon, David. 2000. *Data compression: The complete reference*. New York NY: Springer.

Simon, Herbert A. 1955. On a class of skew distribution functions, *Biometrika* 42(3): 425–440

Stoll, Sabine, Kirsten Abbot-Smith & Elena Lieven. In press. Lexically restricted utterances in Russian, German and English child directed speech. *Cognitive Science.*

Thiessen, Erik D. & Jenny R. Saffran. 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology* 39(4): 706–716.

Tomasello, Michael. 1992. *First verbs: A case study of early grammatical development*. Cambridge: CUP.

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

Tomasello, Michael, Nameera Akhtar, Kelly Dodson & Laura Rekau. 1997. Differential productivity in young children's use of nouns and verbs. *Journal of Child Language* 24(2): 373–387.

Tomasello, Michael & Daniel Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language* 31(1): 101–121.

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

Zipf, George K. 1949. *Human behaviour and the principle of least-effort*. Cambridge MA: Addison-Wesley

Ziv, Jacob & Abraham Lempel.1978. Compression of individual sequences via variable length coding. *IEEE Trans. Inform. Theory* 24: 530–53.

Formulaic language from a learner perspective

What the learner needs to know

Britt Erman
Department of English, Stockholm University

Abstract

The formulaic language in focus in the present paper is collocations. The 'intrinsic' as opposed to 'extrinsic' features of collocations related to Frame Semantics and Lexical Functions are proposed to best reflect their unit-hood status. The paper primarily discusses the lexical status and identification of collocations from different theoretical frameworks, and also reports on a study examining the collocations in English essays written by native and non-native writers. The results show that the non-native group (English students at Stockholm University) have a relatively good command of collocations, but also that their collocational range is reduced and that non-target collocations do occur. The paper concludes with a review of the implications for foreign language teaching more generally.

1. Introduction

What ensures a text's flow and readability? A partial answer to be developed in this paper[1] is that, apart from correct grammar and overall coherence, it is the text's multiword patterns, and expressions, all naturally intertwined and feeding into one another.[2] The multiword expressions in focus are collocations here defined as native speakers' preferred combinations of words. The vast majority of multiword expressions are made up of collocations. However, collocations usually go unnoticed; only when inadequate, erroneous, or otherwise marked collocations disrupt our reading, such as can happen in texts written by L2 writers, do we become aware of their existence and their importance for fluency. The main argument of this paper is that the lack of fluency as evidenced in L2 writings is due to lack of adequate collocations. In other words, collocations are a major stumbling block for L2 writers. It is argued furthermore that the meanings of component parts of collocations are fused, although on the surface the parts look computed and separable. Indeed, many of the collocations identified in this paper have unitary meanings, which suggests that they are formulaic, and, presumably, easily retrieved.

The first and most comprehensive part of the paper discusses the lexical status of a selected set of collocations. With an eye to evaluating what views of collocations would best serve the learner, some current definitions, in particular those adhered to by phraseologists, are brought to light. However, few of the definitions or criteria take a learner perspective. Therefore, in order to meet the needs

1. I would like to thank the anonymous reviewer for valuable comments. Any errors and omissions are solely mine.

2. This research was made possible thanks to a generous grant from the Swedish Research Council.

of the learner, I propose an alternative view of collocations that takes into account the collocation's inherent, or intrinsic, features, i.e., its meaning and function in the context of situation, or frame, to which it would typically belong. Extrinsic factors by contrast, such as restricted choice of members, which constitute the very core of the phraseological framework, are discarded (for a detailed discussion of 'intrinsic' and 'extrinsic', see Poulsen 2005). In fact, limiting collocations to 'restricted' ones would leave out a sequence like *blow the trumpet*, and the majority of similar multiword expressions, on the grounds that there is no restriction on the nominal object the verb *blow* can take and that neither member is used in a specialized or figurative sense (cf. Howarth 1998b). In the approach adopted here a sequence like *blow the trumpet* can only be understood and correctly used against a background frame involving a special instrument, a special technique/activity, a special sound, etc. and is thus considered as a unit with unitary meaning (cf. Fillmore 1985a: 229). On grounds that will become clear in the discussion to follow, this and similar sequences are classified as collocations, although they are perfectly transparent, and, provided we are familiar with the appropriate frames, easily grasped. Therefore, they are multiword expressions that the learner should be encouraged to learn as form-meaning mappings (Ellis 1996) or meaning units.

The second part of the paper presents and evaluates some results from a study of a selected subset of collocations and their lexical status in essays written by Swedish university students of English and a group of native speakers for control. Altogether 30 students took part in the investigation: 15 first-year university students studying at the English department, Stockholm University in 2003; and 15 native speakers of English from various parts of the English-speaking world, who were exchange students at Stockholm University during 2003.

As expected the non-native group used considerably fewer, less varied and sometimes erroneous collocations as compared to the native group. In my discussion of the status of collocations I will largely draw on examples from my own data. The written material thus amounted to 30 essays comprising 8,200 words, i.e., around 4,000 words from each group. Both groups wrote an essay on the same topic, viz. *Is it true that only rich countries can afford to worry about the environment*, which was thought to be general and topical enough for anyone to have views on. The paper concludes with a review of the implications for foreign language teaching more generally.

2. Formulaic language – some voices

John Sinclair's idiom principle (1991: 110) implying that we store a large number of complex items that constitute single choices, although they appear to be the outcome of item-by-item choices neatly sums up ideas about mental storage voiced by

Bolinger as early as the mid-seventies, viz.: "Speakers do at least as much remembering as they do putting together" (1976: 2).

The idea that formulaic language is the first choice of processing language is also expressed by Wray: " [I]t is the accessing of large prefabricated chunks, and not the formulation and analysis of novel strings, that predominates in normal language processing" (2002: 101). Or put differently, language users take what they have stored holistically and analyze only when there is a need to do so (cf. Wray's "needs-only-analysis" 2002: 130–32).

Pawley & Syder (1983) view language production somewhat differently. According to them there is no necessary link between grammaticality and naturalness of expression, meaning that just because a sequence is grammatically correct this does not imply that it is also nativelike. Or, put differently, just because words *can* combine this does not mean that they *do* combine. In fact, if speakers should exercise the creative potential of syntactically correct sentences, they would probably not be judged as exhibiting a nativelike control of the language. One reason why the concept of 'naturalness of expression' had been poorly understood before the 1980s may be that linguistics had been, and to some extent still is, much more concerned with creativity than naturalness (Hoey 2005). The bottom line is that there is a huge store of standard ways of referring to standard situations and phenomena in a speech community. This is the very essence of the relation between language and culture.

Somewhat along the same lines Fillmore (2003) introduces the notion of "the innocent-speaker-hearer" (ISH) whose knowledge is confined to words and word-to-word relations and basic grammatical relations. In other words, the ISH is more likely to know the generalized meanings of single words and how to combine them than the unitary meanings of composite structures. The relation between the ISH's knowledge and the knowledge of multiword expressions, I assume, is the same as that between correct grammar and natural expression in Pawley and Syder's wording.

Next we will consider a special kind of multiword expression, namely collocations.

3. Collocations – general

Collocations are a heterogeneous group of multiword expressions and can take many different forms. The example below (which is the very first sentence of a linguistics textbook) is meant to illustrate that, although they differ in form, all the collocations in this example refer to things that most of us recognize as familiar, everyday phenomena, thus cutting across languages and cultures.

> Every human child given a fighting chance by heredity and environment acquires a native language.

Apart from the compounds *human child* and *native language*, we have the following collocations: a 'restricted collocation' (see 4.2.) holding a metaphorical element (*a fighting chance*), a binomial (*heredity and environment*) followed by a frequent Verb+Noun collocation (*acquire a (native) language*). In the present paper *a native language* and *a human child* are not classified as collocations, but as compounds, primarily on the grounds that the premodifiers function as classifying adjectives,[3] and that they are fixed having reached the ultimate stage of lexicalization (Hudson 1998: 156).

Swedish being closely related to English has cognate renderings of all these collocations, and the same is true of the West-European languages I am familiar with. So, in this example all except the Adj+Noun compounds are classified as collocations foremost because they typically have unitary meanings, and, unlike many idioms, they suffer few syntactic constraints, and, unlike compounds, many of them can be lexically varied although in a restricted manner (see also 4.1). Running the above collocations through Google indeed showed some interesting variational patterns: (1) in the collocation *a fighting chance, fighting* could be replaced with *sporting* but the former is clearly preferred (767,000 vs. 159,000); furthermore, the collocation turned out to be regionally determined so that *sporting* is preferred in the UK and Australia, and *fighting* in the US (for a discussion of regionally preferred collocations, see Erman 2007); (2) in *heredity and environment* the word order could be reversed, but this turned out to be considerably less frequent (129,000 vs. 23,000); (3) in the collocation *acquire a native language* the selected verb *acquire* is much less common than *learn* (16,000 vs. 3,190), which can be explained by the specialized usage due to domain, which, in the above example, is linguistics.

The subset of collocations singled out for the present study includes the following two classes: Verb+Noun and Adjective+Noun combinations.

4. Collocations – some definitions

4.1 Collocations and nativelike selection

Many formulaic structures would be better explained as bona fide idioms, routines, and the like, since they are more or less fixed, not only lexically but frequently also syntactically, hence constituting the very core of what would by most presumably

3. Other Adj+Noun combinations classified as compounds in my corpus data include acid rain, civil war, military force, nuclear power, plastic bottles, public transport.

be called idiomaticity. However, there is another probably more important but at the same time more elusive group of formulaic structures, notably collocations. This group is more difficult to describe but contributes at least as much to native speaker idiomaticity. Collocations are here assumed to be the result of nativelike selection of expression by native speakers to match specific situations and when talking about specific topics, and are a key factor in the mastering of a language. As the name suggests speakers select certain members out of a set. What sets collocations apart from idioms is that many, as was demonstrated above, allow members to be varied, frequently depending on pragmatic factors and the situation at hand. Furthermore, they suffer few syntactic constraints. As a result, the set of possible collocates in collocations is larger than the set of possible members in fixed (or semi-fixed) structures. However, it is my contention that idioms, routines and collocations alike are stored whole and/or easily retrieved.

Pawley & Syder (1983) may have been the first linguists to call attention to questions concerned with what is nativelike in a language. In an attempt to do this they make a distinction between nativelike fluency and nativelike selection. Nativelike selection is defined as "the ability of the native speaker routinely to convey his meaning by an expression that is not only grammatical but also nativelike" 1983: 191). In other words, nativelike selection, at work in the production of collocations, ensures nativelike fluency.

4.2 Frequency-based definitions

Yorio along with a number of linguists defines collocations as "the habitual co-occurrence of individual lexical items". They are combinations of words "that a native speaker would use without thinking much, without an active search for the right word" because they are everyday word combinations (Yorio 1989: 67).

Close to Yorio's definition are definitions suggested by corpus linguists such as: Hoey "[c]ollocation has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (1991: 6–7), Sinclair "[c]ollocation is the occurrence of two or more words within a short space of each other in a text" (1991: 170), Stubbs "[collocation] is a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text" (2001: 24), and Lewis "[c]ollocation is the way in which words co-occur in natural text in statistically significant ways" (2000: 132).

4.3 The phraseologist's view

At one end, then, there is the view that collocations are node-collocate pairs, which are non-directional and probabilistic and which ignore syntax. We have

seen that corpus linguists tend to subscribe to this view. At the other end there is the phraseologist's view that at least one member of the collocation has to be restricted (Howarth 1998b). This view is close to that of proponents of generative linguistics, who regard composite structures as deviations from a standard of full compositionality (Poulsen 2005: 55), i.e., composite structures are not generated from component structures in a systematic and predictable way. In order to sort out composite structures from structures generated by the system their members had to be defined in terms of criterial features. One of the main criterial features for a collocation to be categorized as a phraseological unit is that at least one of its members should be selected from a restricted set (hence the name 'restricted collocation', cf. Howarth 1998b; Cowie 1998b; Nesselhauf 2003) and be used in a specialized, usually figurative, sense. In other words, sequences like *go to seminars/school/work/hospital* would not be assigned phraseological status, on the grounds that, although institutionalized, they contain no figurative element, in contrast to e.g., *go a long way*, which does (in fact, both the verb and the noun phrase are used figuratively). Furthermore, in order for a collocation to be categorized as such it should be unpredictable and/or problematic. In my view it is precisely the collocation's predictability that ensures conventionalization, and vice versa; and, the more conventionalized the combination the stronger the collocational strength. The predictability of conventional expressions not only facilitates encoding and decoding (Poulsen 2005: 77), but also, I would like to add, fluency (cf. Wray 2002).

In this paper the phraseologist's way of categorizing collocations is challenged foremost because it does not take into account the social, psychological and cognitive status of collocations, but treats them as independent of speakers, hearers and situations and other pragmatic and contextual factors. Furthermore, if we take a learner perspective, excluding multiword expressions with unitary meanings from the collocation family because they are not metaphorical or restricted will be fooling the learner into believing that they are computed and compositional. Learners are presumably more interested in the meaning of linguistic expression (cf. Ellis 1996) and 'how things are said' than, e.g., whether or not an expression is metaphorical.

4.4 Collocations in Mel'čuk's framework

Among the first to take a more systematic grip on collocations was Igor Mel'čuk (1996, 1998). According to Mel'čuk the vast majority of what he calls 'phrasemes' are made up of collocations. In fact, he goes so far as to say that phrasemes are "the numerically predominant lexical unit" and outnumber single words by ten to one (Mel'čuk 1998: 24). This does not sound improbable considering that numerous general lexemes are presumably more frequent in collocations than

in word-to-word combinations (e.g., *take, make, do, good, bad,* etc.). Something similar is expressed by Hoey, who says that every word is primed for collocational use (2005: 8).

Strong as these claims may be there is as yet no evidence to the contrary.

4.5 The learner in focus

None of the above definitions and categorizations, with the exception of Mel'čuk (see Section 5.4), serves the needs of the L2 learner. The effect of frequency (cf. Ellis 2002) presumed in the corpus linguist's definition, will normally be marginal, because frequency presupposes exposure, and unless the learner is in continuous contact with the L2 environment, this would be limited in terms of breadth as well as depth. Naturally, the L2 learner could learn through exposure also in a formal, pedagogical environment, but it takes more conscious effort on the part of the learner, and, above all, increased awareness of the nature and abundance of multiword expressions on the part of textbook writers. There are indications that even in large numbers of texts that learners are exposed to they are not likely to encounter repeatedly even fairly frequent words and collocations (Lewis 2008). Furthermore, frequency alone does not ensure unitary meaning, which is of primary concern for the learner.

If the L2 learner's storage of multiword expressions were limited to phraseological units as defined in phraseologists' framework, she would not be able to attain nativelike fluency. As pointed out in the introduction, limiting collocations to 'restricted' or problematic ones would leave out the majority of multiword expressions in any spoken or written text. Furthermore, the 'problematic' multiword expressions for learners are the easy, transparent ones, because they tend to go unnoticed and hence unlearnt.

Fox (2006) and Macqueen (2006) view collocations exclusively from the learner's perspective. Both agree that idiosyncratic language in general, and collocations in particular, cause major problems for learners. Fox emphasizes that whenever even good learners speak or write English the effect is slightly odd, and, the problem is often one of collocation. Macqueen makes a distinction between what can be remedied in learner production (what she calls 'treatable problems', e.g., verb-subject agreement), and what cannot, or at least not without difficulty ('untreatable problems'), which include idiosyncrasy and in particular collocations. Like Fox, Macqueen emphasizes that teaching the vocabulary and basic grammar of an L2 is not nearly enough for targetlike language production, although grammatical errors and obvious vocabulary mistakes are easily spotted and put right (2006). However, neither provides a definition nor criteria for the recognition of collocations that would benefit the learner.

The suggestions in the literature on formulaic language and collocations that will be developed in the following sections and referred back to throughout the article can be summarized as follows:

1. Only a selection of all the potential grammatical units in a language are nativelike.
2. The principle of idiom is the default principle of language processing and not an alternate one.
3. Collocations are a major challenge for learners.

Next follows a proposal for an alternative view of collocations taking into account psychological, social and cultural aspects.

5. An alternative view of collocations

The examples in the main originate from the corpora described in Section 1 above.

5.1 Psychological, social and cultural aspects of collocations

Collocations in this paper refer to composite structures with unitary meanings. They are conventionalized ways of referring to everyday situations preferred by native speakers to other equivalent combinations that could have been selected had there been no conventionalization (cf. Erman & Warren 2000: 31). Rather than define collocations in terms of habitual co-occurrence of words, or specialized/figurative senses, which, as we have seen, are two common definitions, it is argued here that they are above all socially, psychologically and culturally motivated, reflecting language users' experience as social beings at large.

Collocations are a heterogeneous group of multiword expressions, and can make up, or be contained in, all phrase classes (VP, NP, AdjP, etc.). Here three main groups of Verb+Noun and Adjective+Noun collocations are recognized. The first group includes collocations typically serving 'Lexical Functions' (see Section 5.4) including verbal ones, 'support verbs' (e.g., *wreak havoc*) and "fulfilment verbs' (e.g., *burn fossil fuels)*, and adjectival ones (e.g., *appropriate measures)*. The second group includes expressions denoting some specific state, condition, property or activity which is typically socio-culturally motivated, thus invoked by frames, which may be institutions (e.g., *go to seminars, write a check),* or pertaining to social life (e.g., *bright future, entertain friends),* or be the result of democratic processes (e.g., *a free country, a parliamentary debate)*.

Members in these two groups have lexical status, i.e., have specific unitary meanings just like single words and are presumably stored holistically, or at least

easily retrieved, one member calling up the next through associative networks. In fact, pausing within such units is rare (Erman 2007). Furthermore, although a majority of the collocations are transparent in that on the surface the meanings of the parts appear to be intact, the combination gives rise to new meanings (cf. Warren 2005). In other words, appropriate usage usually amounts to more than can be derived from the generalized meanings of the component parts. Although suffering few syntactic constraints, quite a few of them are sensitive to colligational patterns as evidenced in corpora (cf. e.g., Hoey 2005).

Collocations in a third and last group have not (yet) reached lexical status, but are combinations of words induced by specific topics and associated frames. They are presumably easily activated (since frequent) and, above all, expected (given the overall frame, topic and context). Indeed, a text that is short on topically relevant collocations is often impoverished in content, and, in an educational setting, is assigned a lower grade. Topically motivated collocations from essays on environmental issues include *global problem, environmental awareness, environmentally friendly, protect the environment,* etc. (see Section 6).

The first two groups, the Lexical Function group and the Socio-Cultural group, are similar in many ways. Most importantly, they share the characteristic that neither member of the collocation (i.e., neither 'the keyword' (see Section 5.2) nor the collocate) can be 'interrogated', i.e., the focus of a question (for examples see Section 5.3), which suggests that their meanings have fused. This means that we don't need evidence based on frequency from searches in corpora to ascertain their lexical status.

The sections to follow will discuss fusion of meaning (5.3) and the three groups of collocations sketched out above in more detail (5.4, 5.5, 5.6), but first a couple of words about the notion of a keyword.

5.2 The notion of a keyword

There are two ways in which we can view collocations. Either we see them as language internal phenomena, a way which has received its most specific application in Mel'čuk's theory of Lexical Functions. Or, we see them as interplay between language and the external world, which has got its most accurate application in Frame Semantics. Common to these perspectives is the notion of a keyword. Indeed, all three groups of collocations sketched out above share the characteristic that there is a keyword to which collocates are linked in specific ways.

The keyword is assumed to be the collocation's main meaning-bearing element (Mel'čuk 1996: 39). In the selected categories (Verb+Noun and Adjective+Noun) it is the noun that has this role. This is not so strange considering that nouns express things or phenomena that have clear reference, typically constituting the direction/goal of the actions denoted by the verb (as objects of verbs in Verb+Noun

combinations), and their referents being assigned properties or being evaluated through an adjective (typically in Adjective+Noun combinations). Nominal keywords can be anything from high-content, specific words like *penicillin,* which have a small collocational range and thus few collocates, and rather more general, low-content words, such as *situation,* which have many potential collocates (cf. Woolard 2000: 32–33).

Verbs and adjectives are keywords only in combinations with adverbs (Mel'čuk 1996).

5.3 Collocations and fusion of meaning

At the very core of Mel'čuk's framework of Lexical Functions is the notion that a collocation's keyword is semantically chosen, whereas the collocates linked to the keyword are lexically restricted choices imposed by it. However, if we are interested in the lexical status of combinations of words and memory storage, it does not make sense to draw a line between what is semantically and what is lexically chosen, nor which imposes restrictions on which. True as it may be that the noun is the meaning-bearing element of the construction, typically being the topical one, in this paper it is argued that in the selection process neither component has precedence over the other. It is the combination, i.e., the expression as a whole and its unitary meaning/function that is singled out by the text producer. This is explained by the fact that neither the keyword nor the collocate can be interrogated in any natural way. Asking questions about either would place high demands on contextual factors, such as one of the interlocutors having impaired hearing, or the context being generally noisy. So, interrogating the nominal objects 'advantage', 'test' and 'responsibility' in the collocations *take advantage, take a test*, *take responsibility* (i.e., What did you take?) would sound as pragmatically odd as interrogating the verb (What did you do with advantage? With the test? With responsibility?).

Another factor speaking in favour of the collocation being formulaic and having lexical status is position, as apparent in Adj+Noun collocations. The adjective in such collocations would normally prefer attributive position, as in the collocation *a free country.* For example, provided one would like to bring up this particular feature of one's country in answer to the question: What's your country like? the circumstances in order for the adjective to be used predicatively would be highly constrained. In other words, the answer is more likely to be something like: *It's a free country,* than: *?It's/My country is/free.* Or, in the example *It's a small world,* it would be pragmatically odd to place *small* in predicative position? *The world is small,* although, of course, in both examples predicative position is grammatically correct.

On the basis of this I contend that the collocations in the first two groups to be discussed (i.e., collocations being identified in terms of Lexical Functions and those

motivated through socio-cultural conventions) have lexical status, are formulaic and presumably stored/retrieved whole, although, of course, in actual discourse they are frequently separated abiding by the usual syntactic and pragmatic constraints.

Finally, it should be pointed out that, by unitary meaning is not meant invariable meaning. Meanings of collocations like meanings of single lexemes only arise in texts (cf. Sinclair 1992), embedded as they are in superordinate structures and communicative situations. For example, the meaning of the collocation a *free country* will vary according to the overall topic of the discourse, which can be very specific, e.g., 'freedom of speech', or quite general, e.g., a discussion of different political systems.

5.4 Collocations in terms of Lexical Functions

As mentioned, collocations in Mel'čuk's framework are identified in terms of Lexical Functions. There are 34 syntagmatic and 26 paradigmatic Lexical Functions. Paradigmatic Lexical Functions deal with nomination, i.e., with synonymous or derivational correlates of the keyword, e.g., *vehicle* can be used instead of the keyword car. Syntagmatic Lexical Functions deal with combination, i.e., the value of a syntagmatic Lexical Function is used together with the keyword, e.g., *is skidding* is used together with the keyword *car* (Mel'čuk 1996: 46). Only syntagmatic Lexical Functions give rise to collocations and are of interest here. Given Mel'čuk's extremely intricate and technical framework of Lexical Functions, I will have to limit references to it to cases that are relevant for my discussion. This means that several of the subgroups of Lexical Functions will not be touched on. Furthermore, I will focus on 'Standard Lexical Functions', because they have the most comprehensive coverage. I will largely draw on examples from my own data, but occasionally by way of illustration also on examples from Mel'čuk (1996, 1998).

Lexical Functions are abstract categories that can take a variety of linguistic forms. For instance, the Lexical Function involving the extreme point on a scale ('Magn' in Mel'čuk's framework) can be lexically expressed in many ways: as 'stark' in *stark naked,* as 'infinite' *in infinite patience,* as 'highly' in *highly appreciated,* and 'cutthroat' as in *cutthroat competition.* As mentioned, according to Mel'čuk a collocation's keyword is semantically chosen (i.e., a speaker using these collocations in actual discourse has chosen to bring up certain aspects of 'naked(ness)', 'patience', 'appreciation' and 'competition', in our examples). The collocates ('stark', 'infinite', 'highly' and 'cutthroat') linked to the keyword are lexically dependent choices imposed by it (but see 5.3).

Standard Lexical Functions can be adjectival/adverbial, verbal, nominal and prepositional and linguistically expressed in a variety of ways. Only the adjectival and verbal Lexical Functions are relevant for the present discussion. Verbal Lexical Functions include 'support' and 'fulfilment' verb constructions ('Oper' and 'Real', respectively, in Mel'čuk's framework).

5.4.1 *Verbal Lexical Functions*

Support verbs (+ Noun as Object): In contrast to fulfilment verbs, support verbs could be said to be schematic in that their function side is more important than their content side (cf. Paradis' 2001 discussion of these two sides in connection with the construal of adverbs of degree and adjectives). This fact explains why they are sometimes referred to as 'semi-auxiliaries', or described as delexicalized. The nominal keyword is usually a deverbal abstract noun (e.g., 'attempt' as in *make an attempt)*, but not always (e.g., *make an effort)*. Support verbs can be quite general, high-frequency verbs (e.g., *have a shower, take measures, do a favour, make a(n) effort/attempt/mistake,* etc.), but also quite specific (e.g., *launch an appeal, commit suicide, wreak havoc* etc.). Some fall somewhere in between, such as *pose a question, pay attention* and *say a prayer.*

Lexical Functions can combine to make 'Complex Lexical Functions'. For example, there are Lexical Functions that refer to different phases of a process, notably beginning, continuation and end. These can combine with the Lexical Function support to produce e.g., *acquire a habit* (support + beginning), *maintain a habit* (support + continuation), and *drop a habit* (support + end). Sometimes the verb can be stylistically varied (e.g., *kick/get rid of/get out of/break/ a habit*). And, again, there may be colligational constraints, e.g., *they made a decision*, not **they took a decision,* but in the passive 'take' is perfectly all right, *a decision was taken* (although with a slight difference in meaning). The main thing for our purposes is that they are all collocations, which we would want the L2 speaker to learn and store whole, just like the L1 speaker.

Examples from the corpora include, from the native corpus: *implement a policy, set limits, put emphasis (on),* and from the non-native corpus, felicitous ones: *take measures, make an effort, give an example* and less felicitous ones: ?*raise an opinion, *make benefit, *form a question.*

Fulfilment verbs (+ Noun as Object): Fulfilment verbs are recognized by their close semantic relationship with the keyword, which is the nominal object. Their meanings are fused, since they feed into and presuppose one another, by the verb fulfilling the 'requirement' of the keyword. In other words, the verb does with the noun what the verb is supposed to do with the noun (*watch television, consume alcohol, drink beer, drive a bus, listen to music,* etc., etc.). The nominal keyword can be abstract or concrete, and its meaning includes the component that could be rewritten as 'designed to' or 'supposed to' (Mel'čuk 1996: 68). Like some support verb constructions, fulfilment verb constructions can be stylistically varied

(e.g., 'drink' in *drink a couple of beers* could be replaced by 'down' (*down a couple of beers)*. In fact, any synonym will do as long as the Lexical Function is intact.

5.4.2 *Adjectival Lexical Functions*

This group encompasses collocations typically fulfilling Lexical Functions involving adjectives, one of which could be rewritten as 'as it should be', or 'as is the norm', i.e., objective qualifiers that could be said to 'truthfully' describe the keyword ('Ver' in Mel'čuk's framework). As with fulfilment verb constructions there is a close relationship between the keyword and its collocate. The meaning of the adjective is contained in the keyword. For example, if we consider the collocation *appropriate measures,* I think we can all agree that *measures* are typically taken to improve some undesirable state of affairs. In order to achieve this improved state of affairs the measures taken should be 'appropriate'; otherwise they aren't doing what they should be doing. This Lexical Function can combine with one implying the opposite to form a Complex Lexical Function, rewritten as 'as it should not be' ('Anti'+'Ver' in Mel'čuk's framework). Obviously, this Complex Lexical Function (as in the example *inappropriate measures)* does not change the combination's collocational status. Examples of these two Lexical Functions from my native and non-native data include: *possible solutions, valid argument, serious damage, harmful pollutants, a good role model, concrete example, unwise decision, illegitimate power,* the last two of which exemplify the opposite of what 'power' and 'decision' should be, viz. 'wise' and 'legitimate', respectively.

5.5 Socio-culturally motivated collocations

Just like collocations in the Lexical Function category socio-culturally motivated collocations are recognized by having unitary meanings. Furthermore, socio-cultural collocations are like verbal and adjectival Lexical Functions in that they frequently hold a normative element, making reference to societal and cultural norms adhered to in the speech community. What distinguishes the two categories is that in collocations here classified as socio-cultural there is no obvious semantic link between the keyword and its collocate. Needless to say, the distinction is not always clear-cut, and some collocations are indeed on the borderline between the two. For example, the collocation *appropriate behaviour* was originally placed in the Socio-Cultural group, since the keyword *behaviour* was not thought to contain a meaning element that could be rewritten as 'appropriate', a requirement that would have to be met for the collocation to qualify as an instantiation of the adjectival Lexical Function 'as it should be' ('Ver' in Mel'čuk 1996). In other words, there was no obvious semantic link between the keyword and the adjective. However, the related verb 'behave' in certain contexts (e.g., in an utterance like 'Now you behave!' with equal stress on all three words, a common admonition from parents to children), could be rewritten as 'behave appropriately' thus holding an evaluative element. Furthermore, since Mel'čuk (1996: 57) included the collocation *excellent behaviour* to illustrate another adjectival Lexical Function ('Bon', which is distinguished from 'Ver' by being a subjective qualifier), *appropriate behaviour* was finally placed in the Lexical Function category.

Socio-cultural collocations represent sets of values that language users have come to agree on through, e.g., institutions and other bodies of authority, on global (see 5.6) as well as local levels. They are by far the largest group, and since they are everywhere in society we don't pay much attention to them. They can only be explained in terms of our experience of the world as moral, social and cultural beings. For someone who does not have the experience presupposed for a successful interpretation of these collocations, they may simply remain uninterpretable. For example, to someone not familiar with academia the deeper implication of a collocation like *go to seminars* would be ignored. Naturally, the same is true of the majority of collocations pertaining to specific domains of professional life.

So how do we notice socio-cultural collocations? The first answer that comes to mind is: We don't! They only become apparent when they are deemed as 'attempts at target collocations' (Lewis 2005), frequently with interference from L1, such as can be found in the writings of the learners of a language. However, although socio-cultural collocations through their very nature are unlistable, the analyst is at the same time a cultural being and member of a language community. Not surprisingly, therefore, once we start thinking about them and set out to look for them, we find them everywhere. Examples of attempts from my non-native data include **left vegetables* (target: *leftover vegetables*), presumably caused by interference from Swedish, where the same concept is a compound; **plant wheat* (target: *grow wheat*), due to over-generalization of the verb *plant*, rather than support from L1.

Many collocations defined as socio-culturally motivated have arisen because there is a need in a community for expressions reflecting (and constituting) the community's moral values. This is most apparent in proverbs (Schmitt & Carter 2004: 9), but also, I argue, in many socio-cultural collocations. In other words, collocations like *early riser, share responsibility, sustainable development,* etc. (valued as 'good' or morally respectable), and *ignore consequences, selfish people, late riser, fatal effects,* etc. (valued as 'bad' or morally questionable) have not come about by chance, but constitute recurrent topical issues in everyday discourse. And so do collocations like *bleak/dark/dire/future,* although they reflect a slightly different concern, viz. what the future may have in store for us.

Other collocations in the Socio-Cultural group include those that are used to refer to specific, recurrent activities belonging to domains linked to the numerous routines of everyday life, some of which require special tools, or equipment.

For example, when thinking about what we normally do with our teeth, with the floor, etc., what immediately come to mind are the collocations pertaining to the frames connected to these keywords. So collocations like *brush teeth, sweep the floor, polish shoes,* etc. are the conventionalized expressions used to refer to these recurrent activities, thus having lexical status with unitary meanings just like single words. These activities change as new tools, circumstances, etc. see the light of day giving rise to new collocations. For example, before the arrival of the tooth brush people presumably 'cleaned' their teeth. In Germany people still 'polish and/or clean their teeth' *(Zähne putzen),* although I know for a fact that they don't use e.g., a cloth to perform this activity today. So, maybe 'clean one's teeth' is still being used in some hidden corners of the English-speaking world, or by older people. In Sweden we used to 'polish the windows' (Sw: *putsa fönster),* but nowadays we 'wash' them (Sw: *tvätta)*; the older generation would still occasionally use the verb *putsa*. The important thing for the present discussion is that regardless of the lexeme used in connection with the keyword to refer to activities of this kind, the collocation is presumably stored as a unit, and hence should be learnt as such. However, the learner should be particularly careful when using these collocations, because, as we have seen, many of them are language specific.

Another set of nominal keywords in the Socio-Cultural group is of a rather different kind, viz., shell nouns. These keywords are inherently abstract and general; in fact, they have no concrete denotation in themselves but derive their meaning from the 'shell content' stated elsewhere in an utterance or sentential string, thus behaving much like pronouns (cf. Schmid 2000: 13ff. for a comprehensive discussion of shell nouns). The shell nouns appearing in my data include *concern, change, consequence(s), contribution, crisis, problem* and *situation.* They turn out to attract certain kinds of collocates, which also tend to be quite general gathering around what I call 'higher-level lexical dimensions', such as MAGNITUDE or GRAVITY ('problem', e.g., *a major/slight/big problem,* or *serious/grave problem,* etc.), TIME ('crisis', e.g., *current/immediate/future crisis/es)*, IMPORTANCE or MAGNITUDE ('contribution', e.g., *important/relevant/significant/substantial contribution,* or *enormous/major/small contribution* etc.).

Next we turn our attention to collocations induced by a specific topic. The topic concerns questions related to the environment, how it should be taken care of and who should do it.

5.6 Collocations in frames induced by topic

All formulaic language including collocations is created in discourse, so what warrants a special group by that name? The main reason is that we get an opportunity to catch potential collocations belonging to specific topic-induced

frames. Furthermore, we get at differences in the native and non-native speakers' selection of expression related to the topic. So, the main question asked of these essays is: What words in connection with issues to do with the environment will appear in the L1 and L2 writers' minds when writing on what has now become a global concern cutting across languages and cultures?

As in the other two groups of collocations that have been discussed thus far, only adjectives and verbs linked to a keyword were included. Unlike the first two groups of collocations where each collocation had its own keyword, the collocations in this group gather around a set of frame-based keywords belonging to, or associated with, the superordinate frame, viz., ENVIRONMENT. A superordinate frame holds subframes, which are here constituted by lower-level components or aspects of the superordinate frame. Typical subframes in the essays were e.g., THE OZONE LAYER and RAIN FORESTS. Examples in the essays of subframes holding various methods and techniques to secure sustainable development for the protection of the environment include (CLEANER) PRODUCTION SYSTEMS and THE RECYCLING INDUSTRY. These in turn supplied links to other conceptually related frames, the main one being that of ECONOMY (including RESOURCES). Frames and subframes called up in the essays thus provided a selected set of keywords, which were deemed to give rise to topic-induced collocations.

The background frame against which the meanings of words (and collocations) in the ENVIRONMENT frame are to be understood is that our environment is under continuous threat. We get news reports every day informing us of the rate at which the environment is deteriorating. The following quotation from the BNC (the British National Corpus) I think neatly sums up the worry we all have regarding how we should take care of the environment in the best way: "Sometimes human intervention is not a question of changing the environment but of seeking to prevent its change" (AMS486). It therefore came as no surprise that the verbs that most frequently occurred together with the keyword *environment* in the BNC tended to involve protection and preservation of status quo rather than change (e.g., *safeguard, control, save, protect, preserve, conserve, save*). Words involving the opposite of protection, viz., destruction, were also frequent in the BNC (e.g., *damage, harm, destroy, pollute, affect,* etc.). When verbs implying some kind of change were used (e.g., *create, enhance, improve, change,* etc.) the combination at least as frequently referred to a local environment of some sort, or 'environment' used in an abstract sense, with the keyword also typically taking the indefinite article, as in this example from the BNC: "The management is therefore concerned with /…/the deployment of resources to create an environment in which learning flourishes" (B23450).

All the topic-induced word combinations in the native and non-native corpora were checked for collocational status in the BNC and in Google. The advantage of comparing word combinations in texts between different periods,

in this case involving a time span of approximately 15 years, is that we get at the extent to which a combination becomes established over time. For example, there were two instances of *global responsibility(ies)* in the BNC (comprising texts collected between 1991 and 1994), to be compared with 9,121,000 occurrences in Google (2 April, 2007); the same picture emerges for many other combinations, e.g., *strong economies* (BNC: 3 vs. Google: 110,000), *environmentally friendly alternatives* (BNC: 2 vs. Google: 65,700) and *eco-friendly alternative(s)* (BNC: 0 vs. Google: 57,600). It is a matter of course that as soon as subject matters, usually problematic ones, start surfacing in discourse, new collocations are born.

Apart from collocations based on frame-bearing nominal keywords found in the essays, such as *protect the environment, damage the ozone layer, strengthen the economy, strong/weak/good economy, recycle glass, natural resources,* etc. there were plenty of collocations involving shell nouns (see 5.5). Since shell nouns by definition are semantically 'empty', they cannot function as frame-bearing keywords in collocations, but this function will be transferred to the adjective, in the essays realized as e.g., *environmental, economic, ecological, global, renewable.* Examples of collocations from the essays based on adjectival keywords include: *environmental laws/issues/matters/dilemmas/concerns/, ecological change, economic problems/ perspective,* etc.

Summing up: Frames supply topic-induced keywords and these in turn give rise to collocations, which, in an L1 environment, are learnt through daily discursive exposure.

Now, how are collocations handled by learners? In the ensuing sections we will compare the three main groups of collocations presented in the previous sections in essays written by native and non-native speakers on the topic mentioned in Section 1 and repeated here: *Is it true that only rich countries can afford to worry about the environment?*

6. Results of the native speaker/English language learner corpus study

The reader is reminded that thirty informants took part in the study, 15 native and 15 non-native writers.

6.1 Aim and procedure

The aim of the study was three-fold: (1) to measure the proportion of collocations out of the total number of Verb+Noun and Adj+Noun combinations in the essays; (2) to compare the learner and native groups as regards choice and proportion

of collocations; (3) to compare the two groups from the point of view of lexical/ collocational variation.

The first step involved extracting all Verb (incl. particle verbs) + Noun and Adj+Noun combinations. Secondly, in cases of uncertainty regarding the collocational status of the combination it was checked in dictionaries, in the BNC and in Google. On the basis of this two main categories of combinations were distinguished, one where the members lexically preferred one another, i.e., classified as collocations, and one where no such preference was apparent, i.e., 'free' combinations. As everybody knows there is no such thing as a free combination, since selectional restrictions, co-text and pragmatics impose constraints on the combinatorial potential of words.

6.2 Hypotheses

The hypotheses were:

1. The idiom principle is the default principle in all language production for learners and native speakers alike.
2. The learners produce fewer collocations overall and exhibit a smaller variety.
3. The learners would be more uncertain, and, in particular, about natural word combinations relating to specific topics.

The first two hypotheses all rest on the assumption that reference to everyday phenomena in society take special forms of linguistic expression that are composite rather than atomistic, and that this is a universal principle. Considering the L2 learner's normal pedagogic environment we hypothesize that learners have fewer collocations and fewer variants overall. Regarding the last hypothesis it is assumed that writing an essay in a foreign language on a topic which to a great extent depends on L2 discursive input makes accessing appropriate topic-induced collocations more difficult.

6.3 Collocations and 'Free' combinations over the N and NN data

It should be pointed out right at the start that quantifying data that are qualitative in nature is a risky business. Therefore, all the results presented below should be seen as suggestive at best.

The results presented in Fig. 1. show that in their choice of Verb+Noun combinations learners and native speakers alike rely on the idiom principle. In fact, the learners show an even higher figure 71.1% (113/159) of the Verb+Noun combinations vs. 69.5% (171/246) for the native group, but the difference is slight, and statistically non-significant. However, if we collapse the number of collocations

for both groups the native speakers answer for 60.2% (171/284) compared to the learners 39.8% (113/284), or one third more. Seen from that perspective the learners clearly underused collocations compared to the native group. Collocations counted per 1000 words of written text again show a striking difference between the two groups (NN= 0.028 vs. N= 0.043).

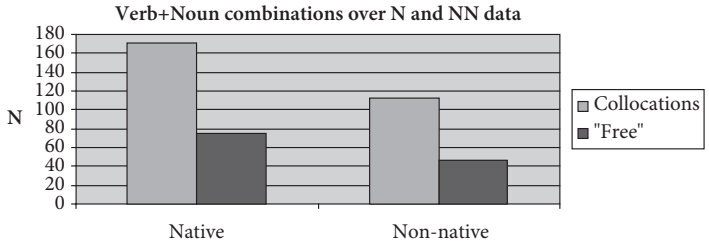


Figure 1. Numbers of occurrence of Verb+Noun collocations and 'free' combinations over N and NN data.

The figures for Adj+Noun combinations showed the same tendency, as shown in Figure 1.

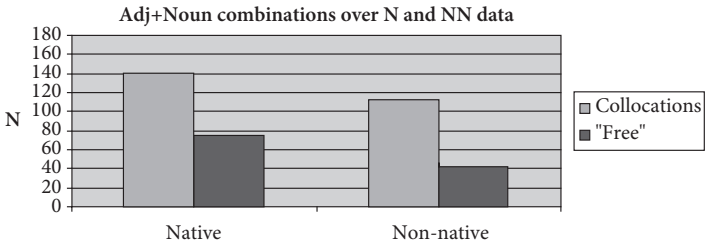


Figure 2. Numbers of occurrence of Adj+Noun collocations and 'free' combinations over N and NN data.

As in the Verb+Noun category, we see that L1 speakers and L2 speakers alike produce more collocations than free combinations, although, as we will see in Figure 3, the learners' selection of collocations does not always match the target language. There is thus a clear tendency among the informants to process language holistically and this may contribute to the learners' high figures for

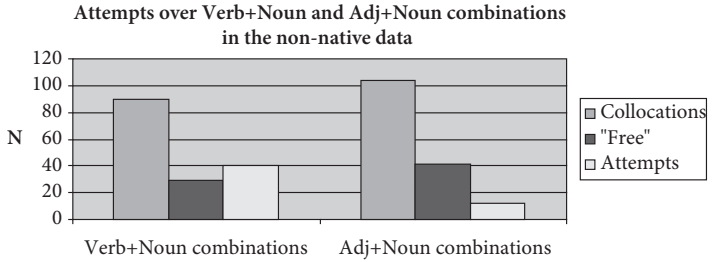


Figure 3. Numbers of occurrence of Attempts in the NN corpus over Collocations and 'Free' in Verb+Noun and Adj+Noun combinations.

collocations also when writing in a foreign/second language. In spite of this, if we collapse the figures we again find that the learners underused collocations 45% of the total number of collocations (114/253) compared to the native speakers 54.9% (139/253).

6.4 Attempts by learners

All the figures in the above diagrams also include less felicitous collocations, or what in this paper have been referred to as 'attempts'. In Figure 3. we correct for this. The figures for attempts are distributed over the categories Collocations and 'Free' in Verb+Noun and Adj+Noun combinations and only concern the non-native group. Although choice of lexis by a native informant in one topic-induced combination (*economical* as in *economical issues*) to some would be marked, it was not classified as an attempt.

However, we can never be certain whether a combination is really an attempt at a formulaic structure, since we are only left with what is in the text and inferring process from product is difficult (Lewis 2005).

Finally, the distribution of the three groups of collocations recognized in the present paper (Lexical Function, Socio-cultural, Topic-induced) for both classes (Verb+Noun and Adj+Noun) among the native and non-native informants can be summarized as follows.

The non-native group scored lower than the native group on all counts, but in particular on topic-induced and socio-cultural collocations. In the Lexical Function group they scored considerably lower on the Adj+Noun collocations, and, among the Verb+Noun collocations support verb constructions were the deficient category. They behaved more like the native speakers on fulfilment verb collocations, which is to be expected considering the close semantic relationship on

which they build. What became particularly apparent in the course of the analysis was that they overused constructions with the verb *have*, such as in *have money, have responsibility, have (financial) resources, have the possibility*, compared to the native group. Examples with *have* from the native data include *have (huge) effects, have a harder time, have the technology, have an obligation, have a role (to play)*. Collocations with *have* found in both groups were *have problems, have money, have resources*. It is to be expected that cognate languages like English and Swedish share many of these collocations, a fact that the learners turned out to put into effect.

7. Discussion of corpus study results

Were our hypotheses met? The answer is clearly in the affirmative.

First of all, there are clear indications that the idiom principle is the default principle in language production for learners and native speakers alike (although the learners' selection of collocations did not always match the target language), which suggests that Sinclair's idiom principle (1991) is the first choice of processing language.

Secondly, the learners produced fewer collocations overall, which supports the second hypothesis to that effect. Furthermore, quite a few of the learners' collocations were classified as attempts, and, in particular, access to topic-induced collocations was limited compared to the native group, which supports the third hypothesis. Finally, as expected, they exhibited a smaller lexical range. This was particularly apparent in the shell noun category.

8. Overall discussion and implications for teaching

One of the main aims of the present paper was to present an approach to collocations that would benefit learners. The approach defended originates from the idea that even transparent multiword expressions and collocations containing no figurative members or members with specialized senses are conventionalized and therefore worth learning as units. It is argued that the focus on phraseological units in the literature has over-shadowed the fact that there is an abundance of multiword expressions that, although they will not pass the phraseological grid, nevertheless have specific, unitary meanings, connected to specific cultural frames, which have to be learnt. In other words, collocations and multiword expressions quite generally go far beyond lexicalized (listable) expressions.

Only when multiword expressions become automatized, i.e., are called up without much reflection, can the learner even hope to attain nativelike fluency.

One way in which multiword expressions can become automatized is through implicit learning (cf. Ellis 2002). How can this be achieved? In order for learners to acquire implicit knowledge we need to maximize opportunities for them to read and use the language, i.e., increase their exposure to the language (cf. Kennedy 2003: 485). Furthermore, teachers should pay more attention to L2 students' lexical errors. The results from the present study indicate that when learners make vocabulary mistakes these are likely to involve multiword expressions, although this is not always brought to their attention. And, although on the surface they look transparent, multiword expressions can only be processed against background knowledge of the external world and of the way in which culture and language are intertwined. This should also be brought to the learners' attention. Indeed, learners are assumed to benefit from a combination of implicit and explicit learning, the latter through raising their awareness of collocations in a pedagogical environment (see Ellis 1997). Finally, more cooperation between teachers, researchers and textbook writers is called for. With increased knowledge of existing corpora and also with how to maximize their use in the production of teaching material, this should not be impossible.

References

Bolinger, Dwight. 1976. Meaning and memory. *Forum Linguisticum* I: 1–14.

Cowie, Anthony P. (Ed.), 1998. *Phraseology. Theory, analysis, and applications.* Oxford: OUP.

Ellis, Nick. 1996. Sequencing in SLA, phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18(1): 91–126.

Ellis, Nick. 1997. Vocabulary acquisition. In *Vocabulary: Description, acquisition, and pedagogy*, N. Schmitt & M. McCarthy (Eds), Cambridge: CUP.

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

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

Erman, Britt. 2007. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics* 12(1): 25–53.

Fillmore, Charles. 1985a. Frames and the semantics of understanding. *Quaderni di Semantica* 6(2): 222–254.

Fillmore, Charles. 2003. A maximalist view of multiword expressions. Paper read at the conference Collocations and idioms: linguistic, computational, and psycholinguistic perspectives, 18–20 September, 2003; Berlin.

Fox, Gwyneth. 2006. Language with attitude. Paper given at the English department, Stockholm University, Stockholm.

Hoey, Michael. 1991. *Patterns of lexis in texts*. Oxford: OUP.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Howarth, Peter. 1998b. Phraseology and second language proficiency. In *Phraseology: Theory, analysis and application,* Anthony P. Cowie (Ed.) , 161–186. Oxford: OUP.

Hudson, Jean. 1998. Perspectives on fixedness: Applied and theoretical. *Lund Studies in English 94.* Lund: Lund University Press.
Kennedy, Graeme. 2003. Amplifier collocations in the British National Corpus: Implications for English language teaching. *Tesol Quarterly* 37(3), 467–87.
Lewis, Margareta. 2005. The elusive formulaic structure: Assessing the adult intermediate learner's idiomaticity, fluency and proficiency. Licentiate thesis, Stockholm University.
Lewis, Margareta. 2008. The Idiom Principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency, and proficiency. Ph.D. dissertation, Stockholm University.
Lewis, Michael. 2000. *Teaching collocation: Further developments in the lexical approach.* Hove: Language Teaching Publications.
Macqueen, Susy. 2006. It just doesn't sound right: An investigation of L2 collocational development. Ph.D. dissertation in progress. Paper given at the School of languages and linguistics, University of Melbourne, Melbourne, Australia.
Mel'čuk, Igor. 1996. Lexical Functions: A tool for the description of lexical relations in the lexicon. In *Lexical functions in lexicography and natural language processing,* L. Wanner (Ed.), 37–102. Amsterdam/Philadelphia: Benjamins,.
Mel'čuk, Igor. 1998. Collocations and lexical functions. In *Phraseology. Theory, analysis, and applications,* Anthony P. Cowie (Ed.), 23–53. Oxford: Clarendon Press.
Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2): 223–242.
Paradis, Carita. 2001. Adjectives and boundedness. *Cognitive Linguistics* 12(1): 47–65.
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), 191–226. London: Longman.
Poulsen, Sonja. 2005. Collocations as a language resource: A functional and cognitive study in English phraseology. Ph.D. dissertation, Institute of Language and Communication, University of Southern Denmark, Odense, Denmark.
Schmid, Hans-Jörg. 2000. English abstract nouns as conceptual shells: From corpus to cognition. *Topics in English Linguistics* 34. Berlin: Mouton de Gruyter.
Schmitt, Norbert & Ronald Carter. 2004. Formulaic sequences in action: An introduction. In *Formulaic sequences. Acquisition, processing and use* [Language Learning & Language Teaching 9],N. Schmitt (Ed.), Amsterdam: John Benjamins.
Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: OUP.
Sinclair, John. 1992. Trust the text. In *Advances in systemic linguistics*, M. Davies & L. Ravelli (Eds), 12–25. London: Pinter.
Stubbs, Michael. 2001. *Words and phrases.* Oxford: Blackwell.
Warren, Beatrice. 2005. A model of idiomaticity. *Nordic Journal of English Studies* 4(1): 35–54.
Woolard, George. 2000. Collocation – encouraging learner independence. In *Teaching collocations: Further development in the Lexical Approach,* M. Lewis (Ed.), 28–46. Hove: Language Teaching Publications.
Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.
Yorio, Carlos A. 1989. Idiomaticity as an indicator of second language proficiency. In Biblingualism across the lifespan, K. Hyltenstam & L. Obler (Eds), 55–72. Cambridge: CUP.

The acquisition and development of the topic marker *wa* in L1 Japanese

The role of NP-*wa*?[1] in mother-child interaction*

Chigusa Kurumada
Department of Linguistics,
Stanford University

Abstract

This study investigates the acquisition of a particular pattern of the Japanese particle *wa*: a noun phrase followed by the particle *wa* uttered with a questioning intonation (NP-*wa*?). Applying the Usage-Based Approach (Tomasello 2003) to children's utterances, we examine the conversations of three mother-child pairs and show that NP-*wa*? is used to

*I would like to thank Shoichi Iwasaki, Shigeru Sakahara, Chie Sakuta and anonymous reviewers for their helpful comments on earlier version of this paper.

1. As outlined in the following discussion, non-phrasal elements can occasionally precede the particle wa (e.g., *Arigatoo wa* ? 'Thank you WA?') However, I use the terminology NP-*wa*? for all examples in this present paper. This is discussed further in Section 3.

draw attention to a referent and to prompt the addressee to comment on the referent. Through their interaction, mother-child pairs often co-construct a complete topic-comment structure, the formulation of which helps children learn new vocabulary and constructions. Analyses suggest that the use of NP-*wa*? creates an optimal context for children to understand grammatical inputs and communicative intentions.

1. Introduction

Applying traditional frameworks for describing adult language to children's utterances taken out of context can be problematic. Children's utterances typically consist of different-sized units deviating from the typical sentences of adults which are usually composed of both a subject and a predicate (Peters 1983). As a result, deciphering the meaning of children's utterances rests primarily on an observer's guess, based on the communicative context and other background knowledge, which are presumably shared by the conversational partners. In opposition to those who view a reliance on contextual cues as evidence of a child's incomplete mastery of language, this study demonstrates that Japanese children competently apply a reduced form of linguistic structure containing the topic particle *wa*, matched with a discourse function to invite a relevant 'guess' from adults. This use of the particle *wa*, first introduced by adults in everyday interactions, is learned by Japanese children through their attempts to communicate with others through joint attention.

Previous studies have examined Japanese children's use of the particle *wa* by comparing it to adult grammar. The particle *wa* has traditionally been described as a marker for a topic NP in a sentence, as illustrated in (1).

(1) *John* *wa* *gakusei* *desu*. 'Speaking of John, (he) is a student.'
John TOP[2] student COP

Since the concept of topic is primarily a discourse notion, the issue of *wa* has been best addressed by researchers who adopt a functional approach to grammar and discourse (Fry 2003; Hinds, Maynard & Iwasaki 1987; Kuno 1973; Noda 1996; Shibatani 1990 among others). Kuno (1973) proposed that the particle *wa* can express different meanings depending on the linguistic and non-linguistic context in which it is used. The meanings include: topic marking, contrastive marking, focus marking, and a marking of quantitative limitation. The multifunctionality of *wa* renders it difficult to provide a clear account of its uses, making the usage of *wa* a major research topic in the field of Japanese linguistics and Japanese-language education.

Based on these findings, many observational studies of the particle *wa* in child language proposed a priori categorizations of *wa* as adverbial particles or topic particles. These studies have found that *wa* appears in children's speech at around eighteen- to twenty-six months of age (Hirakawa 2004; Kuriyama 2001; Nagano 1959; Okubo 1967; Yokoyama 1997). However, as these studies focused primarily on the order of acquisition of different kinds of particles, they did not delve into the question of how children use the particle *wa* in communicative contexts. Although experimental approaches have also been taken to investigate the developmental complexity of the different uses and meanings of *wa* (Hatano 1979; Tahara & Ito 1985), many questions remain regarding the context in which *wa* is used, the role of parental input, and the functions for which children use *wa* in an actual discourse context.

To answer these questions, Clancy (1985: 494) proposes that Japanese children begin using *wa* in the form of 'N *wa*?', as illustrated in (2). She observes that this form of *wa* is used to inquire about the location of an absent entity, or to pose a question about whether a person is to be included in the distribution of food, toys and other things.[3]

(2) *mama* *wa*? 'Where's/What about mommy?'
mommy WA

The use of *wa* with a questioning intonation at the end of an utterance has been investigated by Takagi (2001) from the perspective of conversational analysis. Takagi refers to this linguistic structure as a '*wa*-ending turn' and posits that there are recurrent patterns present in children's use of this question form. In response to children's *wa*-ending turns, adult participants often provide an expanded turn consisting of the nominal phrase initially introduced by the child's *wa*-ending turn, and a predicate which offers some clarification or explanation regarding the referent (131). Children's use of *wa*-ending turns thus elicits a mother's response to their queries, enabling children to initiate conversational turns despite their limited linguistic knowledge and skill.

2. The following abbreviations are used in this paper:

ASP	Aspect	NEG	Negation
TEL	Copular	POL	Politeness marker
FP	Final particle	PAST	Past tense
GEN	Genitive	TOP	Topic
LOC	Locative	QT	Quotation marker

3. Clancy (1996) points out that Korean mother-child dyads make the same formulaic use of the topic-particle in their pointing and labeling routines.

It has also been demonstrated that a child's early use of *wa* in declarative sentences is confined to the case of answering their mother's questions, formulated with *wa* (Clancy 1985), as shown in (3).

(3) Mother: *kotchi* *wa* *dare?* 'Who is this?'
this WA WH-who
Child: *kotchi* *wa* *neesan.* 'This is an elder sister.'
this WA older-sister Clancy (1985: 494)

Clancy claims that these question-answer exchanges foster the acquisition of the particle *wa* because the child benefits from the routinized structure of these interactions and builds sentences upon the preceding utterance.

It is widely accepted that such social interactions have a direct impact on children's later language development. Atkinson (1979), Ochs, Schieffelin & Platt (1979), and Scollon (1976) find that young children and mothers often cooperatively construct a proposition throughout their conversational turns, whereby the child lexicalizes one aspect of the situation and the mother responds to it by lexicalizing another aspect (and vice versa). Atkinson points out that English-speaking children and mothers make frequent use of many lexical items that serve an attention-drawing function, such as *there, see, this, that,* and *look*. By using these lexical items, the child and mother direct each other's attention to the appropriate entity. These lexical items provide an opportunity for the addressee to make a statement about the entity to which they are attending.

Researchers who believe that mother-child interactions have a direct impact on children's syntactic development also propose that children formulate their first hypotheses about the nature of grammar through their experience with directing the attention of their conversational partner (Atkinson 1979; Bates & MacWhinney 1979; Clancy 2003). Bates & MacWhinney posit that children come to understand that a pragmatic topic-comment structure prior to the acquisition of syntax. Recent studies applying the Usage-Based approach to language development appear to support this hypothesis. Tomasello (2003), who reviewed much of the linguistic and psychological research, proposes that children learn the value of linguistic symbols through communicative situations in which the child and the mother jointly attend to a particular object. The child understands the meanings of his mother's utterances by interpreting her communicative intentions. He then learns to reproduce the meanings of the mother's utterances by taking her perspective when he wants others to experience a situation in the same way that he did. Tomasello suggests that children come to understand the intersubjective nature of linguistic symbols and develop pragmatic and syntactic abilities (90).

The aforementioned studies show that Japanese children's use of *wa* can be analyzed using a discourse-pragmatic, usage-based, approach to language acquisition.

Specifically, the turn-initiating function of '*wa*-ending turns' must be related to social interactions under-girding children's use of language. However, due to a lack of longitudinal data, it is not clear how a child comes to use the *wa*-ending turn, structurally defined as an NP followed by *wa* uttered with a rising intonation (henceforth, NP-*wa*?[4]). Data from spontaneous conversations are needed to analyze the influence of parental input and the interaction between mother and child, which is expected to have a large effect on a child's speech. Also lacking is a quantitative approach for examining data on the frequency and usage patterns of NP-*wa*?.

To investigate how and the context of discourse in which Japanese children use the particle *wa*, I address two specific issues:

1. For what functions do Japanese children and mothers use NP-*wa*?
2. How do NP-*wa*? utterances affect the overall development of the use of *wa*?

Examining these questions allows us to clarify how mother-child interactions affect the use and the development of NP-*wa*?. In doing so, I argue that mastery of NP-*wa*? opens the way for children to learn new vocabulary and a broader range of linguistic constructions.

2. Method

2.1 Data

The data for this study were drawn from the CHILDES corpus (Oshima-Takane et al. 1998; MacWhinney 2000) and consist of three sets of mother-child conversations collected longitudinally when the children were aged eighteen- to thirty-five-months (Miyata 2004a, b, c). The three male children, Aki, Ryo, and Tai, were living in the Nagoya area, the third biggest metropolitan district in Japan. They were from middle-class families and received Japanese-only input from their parents. Each mother-child dyad was individually video- and audio-recorded during free-play naturalistic interactions in the home.[5] Mothers were instructed to let their children speak as much as possible but otherwise to play and talk with their children as they usually do. The play sessions were held once a week and recorded for approximately one hour in the cases of Aki and Ryo, and forty minutes in

4. Instead of 'N *wa* ?' in Clancy (1985), this paper uses 'NP-*wa* ?' to refer to the structure because a nominal phrasal structure can appear in the slot preceding *wa* as we see in the following discussion.

5. The video- and audio-recorded data, except for Tai's audio data, have not been made available. The analyses of this present study are based on the transcriptions of these data sources.

the case of Tai. For the purposes of this study, two sessions per month were used for analysis.

2.2 Coding

A total of 3054 utterances containing the particle *wa* were found and classified into 4 types. If the utterance contained a predicate, it fell into Category 1, and otherwise, it fell into Category 2 or 3 depending on whether it was uttered with a rising intonation. Category 4 included utterances consisting of NP-*wa*? with extra elements that were not predicates.

Table 1. Categorization of the utterances that contain the particle *wa*

With/Without predicates	With/Without a rising intonation	Category	Example Utterances	
With predicates		1	KORE *wa usagi-chan. Ryo-kun wa yaru kore.*	'This is a bunny.' 'Ryo will do this.'
Without predicates	'NP *wa*?' with a rising intonation	2	KORE *wa ?*	'What's/How about this?'
	Truncated utterance without a rising intonation ('NP *wa*…')	3	KORE *wa…*	'This is…'
Others		4	*hikooki wa buu tte.*	airplaneWA buzz (onomatopoeia) QT 'the airplane buzz…'

The NP-*wa*? utterances were further categorized into the following seven types according to the linguistic elements inserted into the NP slots.

Table 2. Categorization of the NP-*wa*? utterances

Types of NP		Examples	
NP-*wa*? serving as an NP	deictic pronouns	*kore wa*?	'this WA?'
	common nouns	*densha wa*?	'train WA?'
	kinship terms	*papa wa*?	'Daddy WA?'
	proper nouns	*Rei-chan wa*?	'Rei-chan (sister's name) WA?'
NP-*wa*? serving as a speech-act initiator		*Arigatoo wa?*	'Thank you WA?'
NP-*wa*? serving as an adverbial phrase		*kondo wa*?	'this time WA?'
Others		*koo yuu fuu wa*?	'like this WA?'

2.3 Data analysis

All speech produced by the children and their mothers during the recorded sessions was transcribed according to the JCHAT (Japanese) transcription system (Oshima-Takane, MacWhinney, Shirai, Miyata & Naka 1998). CLAN programs (MacWhinney 2000) were used for quantitative analysis of the data.

To investigate how Japanese children begin to use the particle *wa* and what functions NP-*wa*? serves in interactions, I implemented the following steps of analyses. First, the percentage of time that the particle *wa* was used (against all the words used) was obtained for each child and mother in each pair. Second, utterances that included the particle *wa* were categorized into four types as described in Table 1. Third, the percentage of time when NP-*wa*? was used compared with all utterances containing *wa* was calculated for each mother-child pair. Finally, in order to examine the functions of NP-*wa*?, the NP-*wa*? utterances were classified into seven NP types according to the categorization described in Table 2.

3. Results and discussion

3.1 Utterances containing ***wa***

Figures 1–3 show the percentages of *wa* in comparison with all words used by the three mother-child pairs: Aki-Aki's mother, Ryo-Ryo's mother, and Tai-Tai's mother.

The results show that children only began to make recurrent use of *wa* in their speech when they were nineteen to twenty-five-months old. One child, Tai, was ahead of the other two children in his first use of *wa*; he had a command of *wa* when

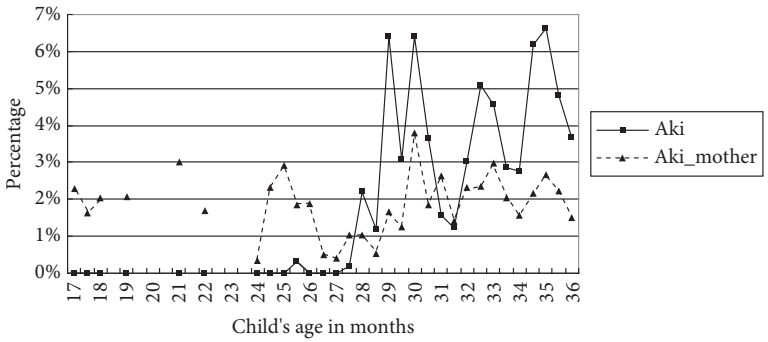


Figure 1. Percentage of time that *wa* was used in comparison with all the words used (Aki and Aki's mother).

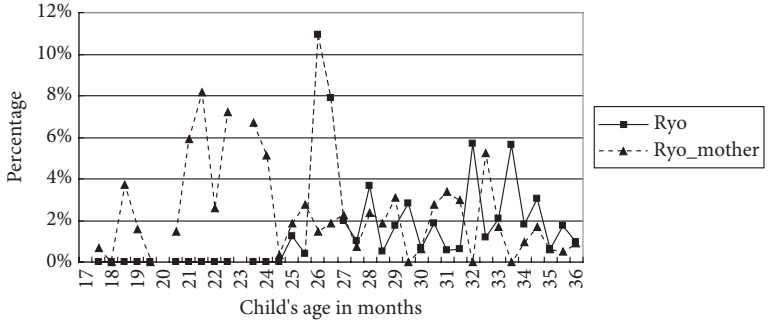


Figure 2. Percentage of time that *wa* was used in comparison with all the words used (Ryo and Ryo's mother).

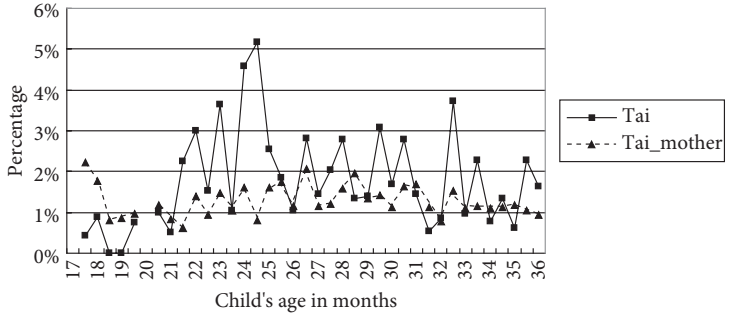


Figure 3. Percentage of time that *wa* was used in comparison with all the words used (Tai and Tai's mother).

the sessions began. The most striking pattern observed was that the mothers used *wa* at a more or less constant rate while the children used *wa* at varying rates. This indicates that the marked increase in the use of *wa* in the children's speech did not result from immediately repeating what their mothers said to them. It is more plausible to assume that the children entered a new developmental stage at that time and began to make extensive use of the particle *wa* for some communicative purpose.

A total of 3054 utterances containing *wa* were categorized into four types based on the classification shown in Table 1 in Section 2.2. The results are illustrated in Figures 4 to 6 below.

The results demonstrate that NP-*wa*? is a particular linguistic formula that both the children and their mothers favored in their speech. When the children were

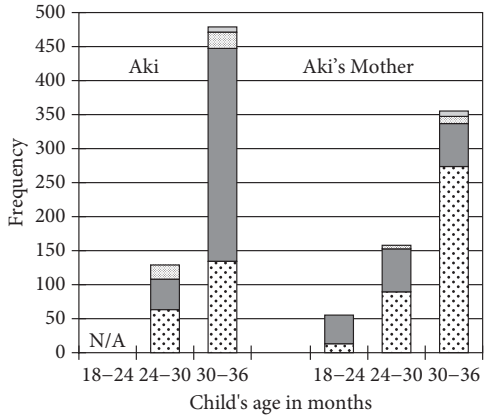


Figure 4. Frequency of the 4 types of utterances containing *wa* (Aki and Aki's mother).

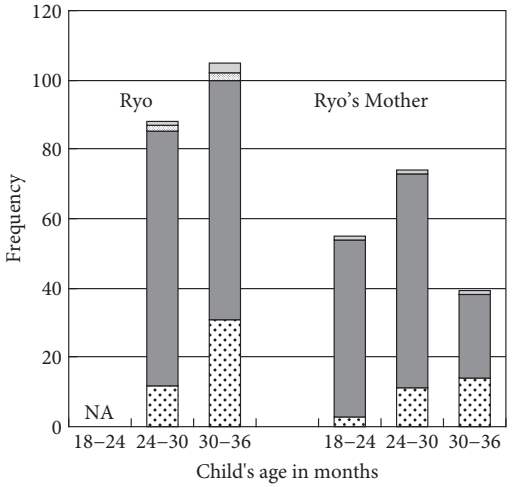


Figure 5. Frequency of the 4 types of utterances containing *wa* (Ryo and Ryo's mother).

younger than thirty-months of age, the mothers made frequent use of the form 'NP-*wa*?' (e.g., *Kore wa*? 'What about this?' *Kuruma wa*? 'Where's (your) car?'). This tendency was much more apparent in the children's speech. Their earliest use of the particle *wa* was largely confined to the form of NP-*wa*?, which remained high in proportion until the children turned thirty-six months of age. In Tai's and his

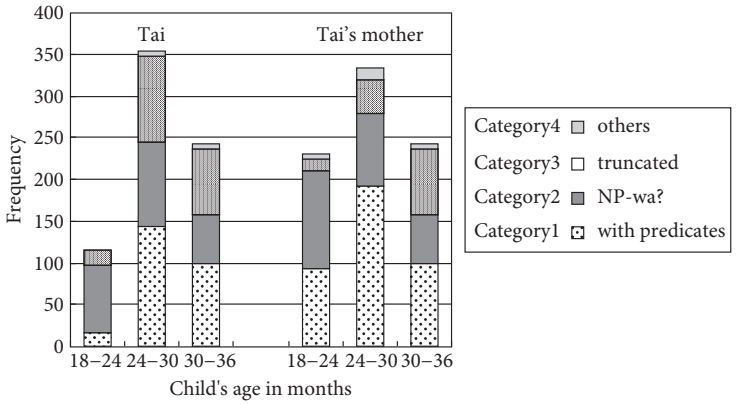


Figure 6. Frequency of the 4 types of utterances containing *wa* (Tai and Tai's mother).

mother's speech, many truncated types (i.e., NP-*wa*.), which lack either overt predicate or rising intonation, also occurred frequently. In sum, for each mother-child pair, a relatively small percentage of utterances containing the particle *wa* accompanied overt predicates. This suggests that the children's and the mothers' use of *wa* cannot be fully explained by examining the distributional patterns of the particle in different argument structures. More attention should be paid to the functions and pragmatic constraints on the use of NP-*wa*?.

The percentages of NP-*wa*? against all utterances containing *wa* are shown in Figures 7–9. Results demonstrate a difference between the children's and the mothers' tendency to use NP-*wa*?. The use of NP-*wa*? was frequent in the mothers'

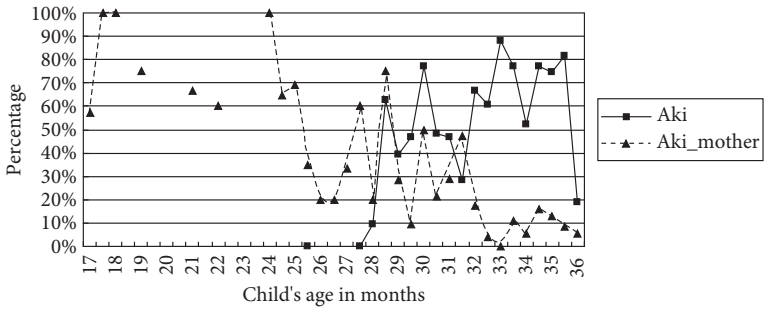


Figure 7. Percentage of time that NP-*wa*? was used against all utterances containing *wa* (Aki and Aki's mother).

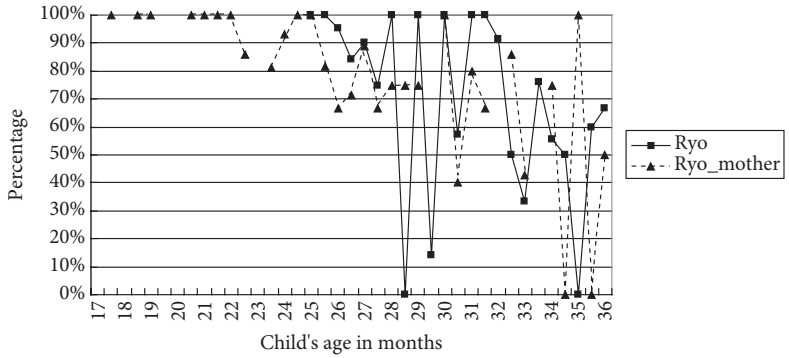


Figure 8. Percentage of time that NP-*wa*? was used against all utterances containing *wa* (Ryo and Ryo's mother).

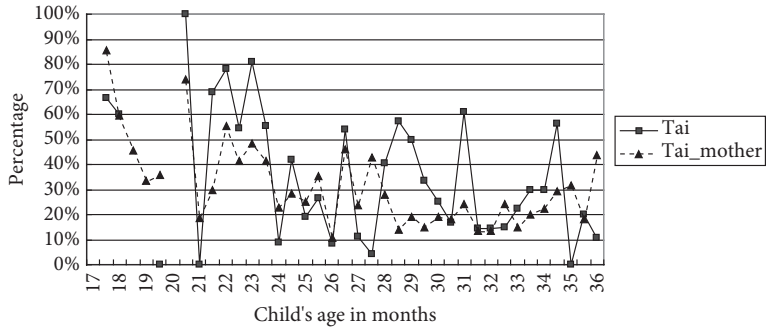


Figure 9. Percentage of time that NP-*wa*? was used against all utterances containing *wa* (Tai and Tai's mother).

speech until the children acquired a good command of the structure. The percentages decreased gradually, however, and remained relatively low in the last six months of the study.

The notable reliance on NP-*wa*? in mothers' speech early in the children's development indicates that the language input that children received tended to be skewed disproportionately. It is reasonable then to assume that some motivation must exist for the mothers to use this pattern. In order to understand the function of NP-*wa*?, we must consider how the mothers used NP-*wa*? in actual communicative contexts.

3.2 NP-*wa*? in mother-child interaction

3.2.1 *How the mothers use NP-*wa*?*

Each mother in the study had a number of routinized interaction patterns with her child wherein she repetitively used NP-*wa*?. Two of the most common uses were the mother's asking about: (1) labels of items and (2) whereabouts of referents.[6] Mothers often used these patterns of NP-*wa*? to ask questions well before the children were able to answer them linguistically. Most of the questions were test questions: mothers deliberately left pauses for the children to respond either verbally or non-verbally and answered if the children did not give a response. The present data show that each of the mothers constantly set up these routines in various care-giving contexts. The excerpt in Example (4) illustrates how Aki's mother used NP-*wa*? to prompt Aki to find an intended referent.

(4) Aki and Aki's mother are looking at the pictures in a book (Aki 21-months-old)

245	Aki:	*ba(su).* bus	'bus'
		(Aki turns pages, sees a picture of a bus)	
246	Aki's mother:	*un* yes	'Yes.'
247	Aki's mother:	*basu doko ni aru* ? bus where LOC exist	'Where is the bus?'
248	Aki	*a.* (Aki points at the bus)	'a'
249	Aki's mother:	*soo, sore basu.* yes that bus	'Yes, it's a bus.'
250	Aki's mother:	***densha wa ?*** train WA	'the train WA?'
251	Aki:	*a.* (Aki points at the train)	'a'
252	Aki's mother:	*un.* yes	'Yes.'
253	Aki's mother:	*sore densha.* that train	'It's a train.'

In this example, Aki and his mother established joint attention through the process of naming objects. The communicative sequence consisted of three steps: the speaker demonstrated his/her awareness of some entity X; the speaker attempted to get the hearer to notice X; the hearer demonstrated that s/he has noticed X (cf. Ochs, Schieffelin & Platt 1979). While it was Aki who initiated this exchange when he found a picture of a bus (in Line 245), the mother reorganized Aki's utterance into a communicative sequence. The mother asked questions including one in the form of NP-*wa*? to draw Aki's attention to objects and Aki answered the questions by pointing and making utterances. The other two mothers also used NP-*wa*? to establish joint attention on an external referent. It may be assumed that these types of NP-*wa*? help the child to understand the relationship between the words he hears and referents to which he and his mother are jointly attending.

Mothers also used NP-*wa*? to draw children's attention to referents and to prompt the children to name them. In Example (5), Tai and his mother are looking at a picture book that contains drawings of various fruits. The mother is attempting to draw Tai's attention to one picture at a time encouraging him to name it.

(5) Tai (15-months-old) and Tai's mother are looking at the pictures in a book

736	Tai's mother:	*kore sakurambo.* this cherry	'This (is a) cherry.'
737	Tai's mother:	*oishii ne, makkana sakurambo.* delicious FP red cherry	'yummy, this red cherry.'
738	Tai:	*ka ka (sui)ka.* watermelon	'watermelon'
739	Tai's mother:	*kore suika .* this watermelon	'This is a watermelon.'
740	Tai's mother:	***Taishoo,***[7] ***kore wa ?*** Taishoo this WA	'Taishoo, this WA?'
741	Tai:	*(ichi)go (ichi)go .* strawberry strawberry	'strawberry, strawberry'
742	Tai's mother:	*ichigo .* strawberry	'strawberry.'

A mother also used NP-*wa*? to prompt her child to repeat words that she has just uttered, as seen in example (6). This pattern differs from searching and labeling routines as in (4) and (5) in that the mothers are not drawing the children's attention to an external object. Instead, they are attempting to attract the children's attention to their own utterances thereby encouraging the children to pronounce the word as the mothers do. Ryo's mother, for instance, produced this type of initiating utterance more than fifty percent of the time that she used NP-*wa*? when Ryo was twenty-four to twenty-eight months old.

6. These two functions of NP-*wa* ? in mothers' speech are closely related to the basic meanings of a Japanese copular sentence containing *wa*. We will come back to this point later in 3.4.

7. Taishoo is the full name of Tai.

(6) Because Ryo's pronunciation is not clear, Ryo's mother tries to teach him how to pronounce the word *tamago* 'egg'. (Ryo 21-months-old)

35	Ryo's mother:	***tamago wa ?***	'eggs WA?'
		egg WA	
36	Ryo:	*tamago*.	'eggs'
		egg	

Some of the NP-*wa*? phrases used for this repetition-initiating function were highly formulaic: their use appeared repeatedly in similar communicative contexts. The most frequent use involved requesting the children to carry out a speech act, such as thanking or apologizing. By using this pattern the mothers modeled utterances that the children then repeated, as shown in the examples (7) and (8).

(7) Aki 27-months-old

171	Aki's mother:	***gomennasai wa***?	'I'm sorry WA?'
		I'm-sorry WA	

(8) Ryo 26-months-old

355	Ryo's mother:	***arigatoogozaimashita wa***?	'Thank you very much WA?'
		thank-you-POL-PAST WA	
356	Ryo:	***arigatoo***.	'Thank you'.
		thank-you	

This is a special case of the use of NP-*wa*? because the slot preceding *wa* can contain non-phrasal linguistic elements such as a clausal structure with a subject and a predicate. However, the present data show that this type of NP-*wa*? is not an exceptional usage, rather it is one of the most frequent usage patterns of NP-*wa*? in Japanese motherese. Clancy (1985) reports that the speech of Japanese caregivers' is characterized by a high frequency of explicit directives that demonstrate how children are supposed to speak and behave in various contexts. NP-*wa*? differs from other linguistic expressions used to control children's behavior in that it is applied to 'remind' rather than directly demonstrate how to do things in a particular way (Burdelski 2006). For example, when a mother says '*Arigatoo wa*? (Thank you WA?)', she relies on the child's pre-existing knowledge about when and how to say 'Thank you' and tries to remind him to say it in the given communicative context. In this respect, it can be safely stated that NP-*wa*? calling for a speech act and NP-*wa*? referring to an outside entity share the same basic attention-getting and response-eliciting function. The mother uses NP-*wa*? to control the child's attention to a referent of the NP, whether in the outside world or within the knowledge of the child, and she prompts him to signal his understanding of the referent by pointing, showing, commenting, or doing the intended speech act. These findings lead us to propose that the mothers are using

NP-*wa*? to keep their children involved in an interaction, partially by controlling their attention and behavior.

The examples (4) through (8) indicate that the mothers' use of NP-*wa*? elicits a variety of responses from the child depending on the communicative context. In other words, through the NP-*wa*? routines the mothers guide their children to discern their interlocutors' communicative intentions. To achieve this, children are required to understand the mothers' intention at two levels. First, the child needs to recognize that his mother wants him to answer her question. This understanding is far more complicated than it appears because it requires an understanding of the embedded structure of communicative intentions such as: [You (Mother) intend for [me (Child) to respond to the utterance [X]]] (cf. Tomasello 2003: 23). Secondly, the child needs to detect what kind of answer the mother expects him to give. Miyata (1992) claims that NP-*wa*? is used as a kind of WH-question in children's speech in that the child can ask for various types of information about the referent. The mothers' use of NP-*wa*? in the present data also has functional similarity to a WH-question, with the content of the WH-word remaining implicit, thereby requiring that the child interpret the mother's communicative intentions. Therefore, in the social-pragmatic view of language acquisition, mothers' frequent use of NP-*wa*? in the early stage of children's language development provides a context wherein the children must infer the communicative intentions of their mothers.

3.2.2 *How children use NP-*wa*?*

When children acquire a good command of *wa* by twenty-five to thirty- months of age, the roles of questioner and responder appear to be reversed. Before the children reach this stage it was primarily the mother who used NP-*wa*? to ask questions. At around thirty months of age the children began to use NP-*wa*? to direct their mothers' attention to a particular object and to elicit a response to their queries.

In sharp contrast to the variety of communicative purposes for which the mothers used NP-*wa*?, the children used this structure for relatively restricted communicative purposes. Figure 10 illustrates the percentages of the types of NPs in NP-*wa*? used by the three children in each six-month period. (The categories are presented in Table 2 in Section 2.2) It clearly shows that the children had a notable tendency to employ demonstrative pronouns (*kore* 'this', *koko* 'here' and *kotchi* 'here/this one'[8]) when they used NP-*wa*?, forty to seventy percent of the

8. Almost all the examples of demonstrative pronouns were either '*KORE*' (this) '*KOKO*'(here) or '*KOTCHI* '(here/this one). There were a few utterances containing '*are*' (that) or '*asoko*' (there).

time. This is remarkable because the percentage of demonstrative pronouns in NP-*wa*? constantly exceeded that of common nouns, which are comprised of a wider variety of items compared with other categories.

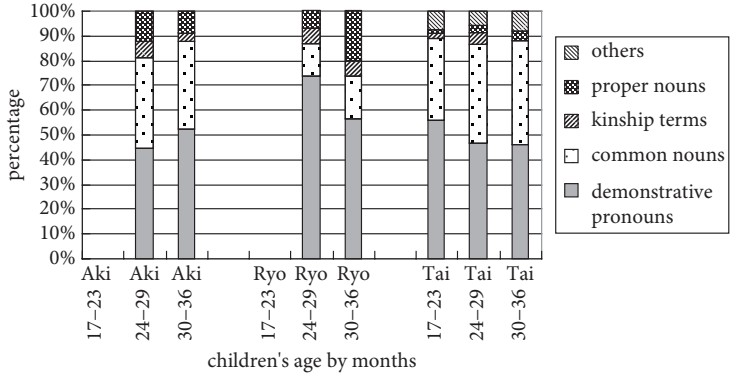


Figure 10. Types of NP in NP-*wa*? utterances.

In sum, at the earliest stages of acquisition, the particle *wa* most frequently appeared as part of a formulaic structure: a deictic NP followed by *wa* pronounced with a rising intonation. This structure is formulaic in the sense that it 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: 1). Only a small variation is allowed in the selection of linguistic items appearing in front of the particle *wa* at this stage in the use of NP-*wa*?, making it difficult for children to generalize the structure as a combination of a variable slot (i.e., an NP) and a pivotal element (i.e., *wa*). Rather, NP-*wa*? containing different types of demonstrative pronouns may be used as concrete and prefabricated expressions with no clear relationship to each other.[9]

The frequent appearance of deictic expressions seems to suggest that children used NP-*wa*? primarily to draw the listeners' attention to an object in the immediate context. The formulaic structure of NP-*wa*? is highly useful in this situation as

9. Hashimoto & Amano (2007) present data which support this observation. They collected data from one Japanese family with a child from the age of eighteen to forty-eight months and extracted any utterances containing the particle *wa*. The data demonstrate that the child used '*kore wa* ?'(a demonstrative pronoun 'this' followed by WA?) 66.7% of the time that she used *wa* when she was twenty-four months old. The percentage, however, plummeted to 31.8% at the age of thirty-six months and 7.4 % at forty-eight months old. Thus, it took awhile before her use of *wa* become generalized across a wide variety of construction types.

the children can use it even when they do not know the name of the object; they avail themselves of demonstrative pronouns to deictically refer to it. The mother's answer can also be minimal, consisting only of the queried predicate as in (9). Alternatively, the mother may expand the child's utterance by replacing it with fuller grammatical structures as illustrated in (10).

(9) Aki and Aki's mother are looking at a picture book (Aki 30-months-old)

681	Aki:	***kore*** ***wa*** **?**	'this WA?'
		this WA	
682	Aki's moher:	*arisan*.	'an ant'
		ant	
683	Aki:	***kore*** ***wa*** **?**	'this WA?'
		this WA	
684	Aki's mother:	*tentoomushi*.	'a ladybug'
		ladybug	

(10) Tai and Tai's mother are looking at a picture book (Tai 28-months-old)

1302	Tai:	kore wa?	'this WA?'
		this WA	
1303	Tai's mother:	*kore wa pawaashoberu*.	'This is a power shovel.'
		this WA power-shovel	

In these cases, the child and mother produced a question and answer, which resulted in the co-construction of a full propositional structure. Grammatically, the NP in an NP-*wa*? becomes an argument for the subsequent predication and thus the combination of NP-*wa*? and the response make a statement. Table 3 illustrates the percentages of different types of mothers' utterances containing the particle *wa* in each six month period. The percentages of NP-*wa*? and percentages of utterances composed of NP-*wa* and a predicate are compared in Figure 11. These data demonstrate that, as the children grew older, the mothers' use of *wa* occurred more frequently in the constructions containing predicates[10] while the use of of NP-*wa*?, which was dominant in the first period (eighteen to twenty-four months), decreased steadily. This increasing variation in the usage of constructions containing *wa* may be due to the fact that the mothers were no longer the only ones asking questions by means of NP-*wa*?. They were now being asked questions by their children and answering them in sentences characterized by varied and complex structures such as (10). Through these mechanisms, we believe that

10. These two functions of NP-*wa* ? in mothers' speech are closely related to the basic meanings of a Japanese copular sentence containing *wa* . We will come back to this point later in 3.4.

children gradually learned how to articulate propositions by producing an utterance composed of both a subject and a predicate.

Table 3. Percentages of different types of the three mothers' utterances containing the particle *wa* (%)

		18–24 months	24–30 months	30–36 months
with predicates (nominal, adjectival, and verbal predicates)		23.5	42.9	51.6
without predicates	NP-*wa*?	72.6	49.8	34.3
	NP-*wa*…	2.0	5.1	11.8
others		1.9	2.2	2.3

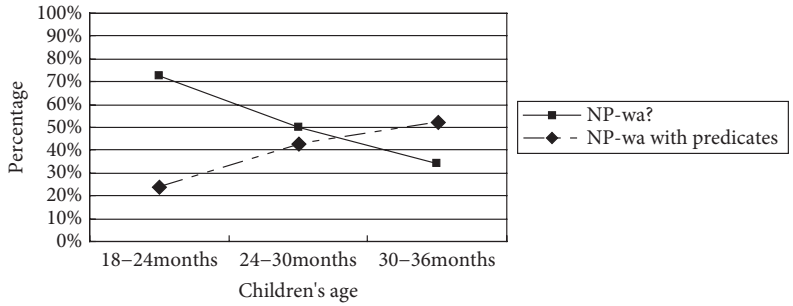


Figure 11. Percentages of NP-*wa*? and utterances composed of NP-*wa* and predicates in the three mothers' total utterances.

To summarize, the children learned that the particular structure containing the particle *wa* serves a communicative function: to get a listener to pay attention to a referent of the NP preceding it and to request a comment about the referent from him/her. Evidence of this strong form-function relationship is seen in the children's creative uses of NP-*wa*? observed in the present data. In (11), Tai placed *wa* after another potential topic marker *mo*, and in (12) Aki placed *wa* after a question word (*dare* 'who'), both of which are unacceptable in adult grammar.

(11) Tai, 26-months-old
1021 Tai: ***kore mo wa?*** 'this, too WA?'
this too WA

(12) Aki, 35-months-old
550 Aki: ***dare wa?*** 'who WA?'
who WA

Table 4. Percentages of different types of the mothers' utterances containing the particle *wa* (%)

		18–24 months			24–30 months			30–36 months		
		Aki's mo*	Ryo's mo	Tai's mo	Aki's mo	Ryo's mo	Tai's mo	Aki's mo	Ryo's mo	Tai's mo
with predicates	Nominal Predicate	5.4	3.6	11.3	22.6	6.8	23.7	25.1	10.3	15.9
	Adjectival Predicate	1.8	0.0	10.9	5.7	2.7	8.7	18.1	7.7	8.4
	Verbal Predicate	17.9	1.8	17.8	28.3	5.4	24.9	33.9	17.9	17.5
without predicates	NP-wa?	73.2	92.7	51.7	39.6	83.8	26.0	17.5	61.5	23.9
	NP-wa…	0	0.0	6.1	3.1	0	12.3	3.4	0	31.9
others		1.8	1.8	2.2	0.6	1.4	4.5	2.0	2.6	2.4

*mo–mother

These creative uses of NP-*wa*? have also been reported in previous research (Nagano 1959; Takagi 2001; Nakayama & Nakayama-Ichihashi 2000). These studies regard them as an erroneous or over-extensive use of *wa*. According to Takagi, Nakayama & Nakayama-Ichihashi (2000) propose that the creative use of NP-*wa*? is an extension of the more fundamental textual functions of the particle *wa* (i.e., thematization of a referent) through the process of grammaticalization. Alternatively, the present paper attempts to provide a unitary account for these seemingly erroneous uses of NP-*wa*? and the uses which involve NPs such as demonstrative pronouns and other common nouns. In both cases, the children may apply NP-*wa*? with the same functional goal in mind: getting the addressee's attention and prompting him/her to respond to them.

Recent studies applying a constructional approach to language development (Brooks &Tomasello 1999; Clark & Kelly 2006; Goldberg 2006; Tomasello 2003; Tomasello & Brooks 1998; among others) have noted that young children are attentive to statistical regularities exhibited by the linguistic elements appearing in a particular grammatical schema. Children can detect commonalities in the way the linguistic items function and in the configuration of these items across different utterances. By observing these statistical regularities they can construct hypotheses regarding abstract syntactic categories. In this light, the ostensibly erroneous uses of NP-*wa*? suggest that the young Japanese children actively test their hypotheses about the linguistic category whose member can appear in the variable slot followed by the particle *wa*. Eventually, the unorthodox uses are expected to disappear naturally as children gain more experience using NP-*wa*?. At the same time, the structure of NP-*wa*?, which arose in children's language as a tightly formulaic expression containing a small number of deictic pronouns, will become more abstract by including many different kinds of linguistic symbols.

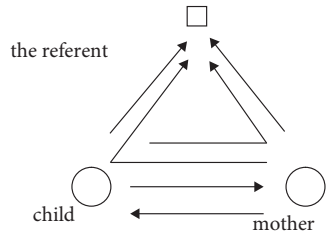
3.3 How the use of NP-***wa***? affects the overall process of language development

3.3.1 *Joint attention and the use of NP-*wa*?*

We have seen that the use of NP-*wa*? is cultivated in the routinized turn-taking patterns of children and mothers. NP-*wa*? was used by the children to draw the addressee's attention and establish a triadic relationship, or a 'joint attentional frame' (Tomasello 2003), consisting of the child, the mother, and the referent of the NP. This situation is illustrated in (13). As proposed by Tomasello, the joint attention frame provides the young child with an optimal context in which to understand grammatical inputs with respect to the child and the caregiver's communicative intentions. NP-*wa*? , in this sense, provides the context for establishing

a framework where a child can learn that the mother is using a piece of language for a particular communicative goal.[11]

(13) The visualized image of joint attention (Tomasello 2003: 29, modified by Kurumada)



My claim is that in the process of Japanese acquisition, NP-*wa*? plays a unique role in the formulation of the triadic relationship. As is proposed in Section 3.2., the mothers used NP-*wa*? with a variety of communicative intentions in mind. To respond to the mother's use of NP-*wa*?, the child has to be aware of his mother's attention to both himself and to the intended referent. As the child becomes accustomed to this communicative structure, he learns to control his mother's attention in the same way by reversing the directions of illocutionary force (Austin 1975). Tomasello calls this process 'role-reversal imitation', in which the child aligns herself with the adult in terms of both the goal and the means for attaining communicative goals (2003: 25–28).

It is highly likely that the unique formal and functional features of NP-*wa*? make the role-reversal imitation easily achievable for Japanese children. As a child repeatedly experiences the question-answer exchange including NP-*wa*?, it is likely that he learns the relationship among: (1) the grammatical features of utterances, (2) the functional roles of utterances, and (3) the roles of speech participants. These three aspects are illustrated in the table below.

11. Boldwin (1991) finds that 16- to 17-month-olds succeed in learning words when the child and adult are both looking at the same object, but they fail when both parties are focusing on different objects. Interactions initiated through using NP-*wa* ? are therefore expected to facilitate the child's lexical learning and syntactic learning through the co-construction of sentence structures.

Table 5. Relationship between NP-*wa*? and the following utterance

	NP-wa?	The following utterance
Grammar	To present the NP as an argument	To provide the predicate with the argument and/or the predicate
Function	To elicit the listener's attention and ask a question	To evidence awareness of the referent and answer the question
Roles of speech participants	The questioner as a topic presenter	The respondent as a comment provider

It is logically assumed that the conventionalized relationship between grammar, the function, and the role of the speaker reduces the cognitive burden on children who must take on different roles in a communicative situation. As a result, the child and the mother can switch between the roles of questioner and respondent. By playing these two roles, the child begins to understand both the mother's role and his own role in the interaction from the same outside perspective (Tomasello 2003: 22) which may help children comprehend the symbolic use of language.

The use of NP-*wa*? is crucial in this process because it not only directs the recipient's attention to an intended referent, but it also obliges her to provide a response. Tanaka (1999) proposed that NP-*wa*? in adult Japanese conversation has 'projectability properties'. That is, the recipient of an NP-*wa*? utterance is pragmatically obliged to infer the underlying topic-comment structure which is projected by NP-*wa*?.[12] Based on this argument, Takagi (2001) posited that NP-*wa*? in mother-child interaction has the same property in that it requires that the interlocutor provide a verbal answer to a query. Consequently, by engaging in a simple question-answer sequence, the mother and the child jointly attend to a third entity and then co-construct a propositional structure involving the shared topic.

Due to its projectability properties, the use of NP-*wa*? enables children to take an active role in the process of collaborative communication. In the present data, about forty to sixty percent of the time that the children used NP-*wa*? (Aki, 63.4%, Ryo, 45.7%, and Tai, 39.6%), the mothers respond to them by either (1) producing queried predicates only, or (2) recasting the utterance by producing both the arguments and the predicates. The use of NP-*wa*? thus makes it possible for children to initiate the joint construction of a single propositional structure, allowing for the expansion of their lexical knowledge and more complex syntactic structures. They are therefore scaffolding their own language learning by drawing adult interlocutors into a referential triangle.

12. "When *wa* is used at the end of a questioning turn, it is always the predicate associated with the introduced NP that is 'projected'"(Takagi: 188). Though particles other than *wa* may share the same function, *wa* obliges the listener more strongly to provide a comment on the introduced NP. In the following excerpt cited from Takagi (185), the object marking particle *o* is used in the form of NP-*o* ?. It only refers to the propositional content of the preceding utterance and has no projecting properties.

A: *jibun toko de udon tsukutteru* '(They) make udon (noodle) at their place.'
B: *oudon o* ? 'Udon O?'

3.3.2 *Development of a sentential structure*

The analysis of the use of NP-*wa*? in mother-child interaction provides further insight into how the dyadic construction of a proposition including *wa* can be internalized into a child's linguistic knowledge for his solo production of the Japanese copula sentence *A wa B da* 'A is B'. The copular sentence has two distinct uses: predication and identification (Sakahara 1996: 262), as seen in examples below. They are not encoded syntactically but distinguished by the relative imbalance of the referentiality assigned to the two NPs. (14)-a serves to express a predication about the subject referent, Tokyo, which conveys higher referenciality than the descriptive noun phrase 'the capital city of Japan' does. (14)-b, on the other hand, achieves the identification of a referent that best fits the description provided by the subject noun phrase.

(14) the two uses of a copular sentence

a. *Tookyoo wa nihon no shuto da.*
Tokyo WA Japan GEN capital COP
'Tokyo is the capital city of Japan.' [predication]

b. *Nihon no shuto wa TOOKYOO da.*
Japan GEN capital WA Tokyo COP
'The capital city of Japan is Tokyo' [identification]

The present analysis shows that, in the routinized act of questioning and answering including NP-*wa*?, a child and a caregiver collaboratively attain the two discourse functions of a copular sentence. As we have seen in 3.2, the mothers frequently asked: (1) labels of items and (2) whereabouts of referents, which in many cases fall into an act of a predication or an act of identification respectively. Children, who have a smaller vocabulary and a stronger inclination to make a deictic reference, tend to initiate a co-construction of the predication type by pointing at the intended referent and asking their mothers for a predication. In either case, a copular sentence represents a pragmatic format wherein a child acquires a new piece of knowledge by actively connecting the referent with a more abstract linguistic notion provided by the non-referential NPs (S.Sakahara, personal communication, December 5, 2007).

Furthermore, *wa* in the predicational type of copular sentence has traditionally been regarded as a 'separating particle' which has the dual functions of (1) separating an entity from the rest of the event or state, and (2) re-uniting the entity and the event or state to make an emphatic judgment (Shibatani 1990: 264). For example, in stating *sono hana wa kiree da*. 'The flower is beautiful.', the speaker extracts the idea of the flower from the situation he wishes to describe instead of depicting the situation as a whole. He can then lexicalize a property of the flower by predicating that it is beautiful. In the present analysis, from an ontogenic point of view, it is observed that an NP followed by *wa* is literally separated from the comment part in children's earliest uses. The recipient then adds a comment, which completes the proposition. The separating and the uniting functions are carried out by two speakers through their conversational turns.

This provides insight into what a Japanese child needs to learn to be able to produce an utterance including *wa* consisting of both an argument and a predicate. That is, he needs to play both roles of separating and uniting (i.e., topicalization and predication), which would typically be conducted by two speakers in earlier stages of linguistic development. Put differently, the child needs to learn that it is only after he identifies an entity and directs the listener's attention to it that he can make a statement about it. The problem is therefore related to his pragmatic ability to assess the informational state of the listener. The child needs to understand what is known and what is regarded as 'certain' (Greenfield 1979) for him and others. At earlier developmental stages the listener's knowledge is not as important to the child as he produces utterances regarding what is new and salient for him. However, when both parties begin producing both topics and comments at the same time, the child faces the problem of respecting the listener's needs and modifying his speech accordingly.

In the present data, there is evidence that children achieve this task by interacting with their mothers. Children often receive explicit cues from their mothers which assure them that their mothers are paying attention to the same object that they are. In the following excerpt (15), Tai and his mother were talking about characters of the cartoon 'Thomas the Tank Engine' and Tai's mother asked how he liked James (one of the tank engines) by means of NP-*wa*?. Tai repeated the question, seemingly to ensure that they were in fact discussing the same topic. He then produced the same NP followed by *wa*, this time without a rising intonation, and a predicate in different turns. Each of Tai's utterances was followed by feedback from his mother.

(15) Tai (31-months-old) and Tai's mother are talking about the characters of 'Thomas the Tank Engine'

747	Tai's mother:	*fuun* ***Jeemusu*** ***wa*** ?	'well, James WA?'
		OK James WA	
748	Tai:	***Jeemusu*** ***wa*** ?	'James WA?'
		James WA	
749	Tai's mother:	*un*.	'Yes.'
		yes	
750	Tai:	*Jeemusu* *wa*.	'James WA'
		James WA	
751	Tai's mother:	*un*.	'Yes.'
		yes	
752	Tai:	*kattenakatta*.	'[I/We] did not buy.'
		Buy-ASP-NEG-PAST	
753	Tai's mother:	*fuun*.	'OK.'
		OK	

The object that the mother referred to by means of NP-*wa*? was not immediately commented on by Tai. The referent was reestablished as a topic in Line 750, which was commented on by the forthcoming predicate '[I/We] didn't buy' in line 752. At this stage, the 'canonical' topic-comment structure with an overt predicate can only be seen across the discourse turns of two speakers (in lines 750 – 752 'James, [I/We] didn't buy.'). The child and the mother are constantly verifying each other's attentional state, which appears to make it easier for the child to assess the mother's knowledge. The topic-comment structure, first realized as a question-answer exchange, is thus gradually internalized in the child's linguistic knowledge while enlisting the aid of adults' utterances and the discourse context.

4. Conclusion

The purpose of this research has been to demonstrate that the use of the Japanese particle *wa* – the multifunctional word whose behavior is not easily explained – comes into being in a particular constructional pattern (i.e., NP-*wa*?) linked to unique mother-child interactive activities. Findings show that the mothers utilized this pattern well before the children were able to respond to the utterance through linguistic means. The primary function of the mother's NP-*wa*? at this stage was to draw the child's attention to a third object and oblige him to respond either verbally or non-verbally. The three mothers in the present study used NP-*wa*? in an attempt to establish a communicative frame in which the mother and the child communicate while attending to the intentional state of the other.

Around their second birthday, the children acquired the command of this specific construction. The rudimentary forms of NP-*wa*? evolved from highly formulaic patterns which included limited types of demonstrative pronouns in the NP slot. Using the formulaic type of NP-*wa*?, the children began to direct their mothers' attention towards an intended referent in a conversational context. In a pragmatic account, this referential act often functions to present a topic and

then invites a linguistic comment from the mother. Thus the children effectively made use of NP-*wa*? to draw the mothers into a co-construction of a proposition thereby generating the input, both lexical and constructional. In sum, the most important contribution of NP-*wa*? to language acquisition is in enabling children to take an active role in the establishment of the referential triangle wherein two speakers attend to each other's intentional state.

The present research also highlighted the role of mother-child interaction in fostering children's ability to understand communicative intentions. Many studies investigating language acquisition from the standpoint of children's socio-cultural development postulate that the ability to understand the intention of others develops around a child's first birthday. However, analysis of the use of NP-*wa*? demonstrates that the intention-reading skill itself is cultivated within a mother-child interaction. At each step, the mother incorporated whatever competencies the child had already acquired to encourage him to assess her awareness of his own intentional state. She was controlling his attention, asking a question, providing an answer if he lacked one, and accepting an answer if he gave one. Once the child became competent in imitating the form and the function of NP-*wa*?, the mother would respond to or recast his queries. The child's mastery of intention-reading is therefore considered to be accomplished through interactions with his mother on the basis of the mother's inferences about the child's competencies.

Future research should investigate how often this kind of co-construction occurs and what kind of arguments and predicates are produced by children and mothers. It is also important to investigate how children's gestures and physical movements are related to the use of NP-*wa*? This line of research will illuminate the possibility of analyzing the interface between children's socio-pragmatic abilities and language development, and how children's language arises from fundamental communicative and referential strategies.[8]

References

Atkinson, Martin. 1979. Prerequisites for reference. In *Developmental pragmatics*, E. Ochs & B. Schieffelin (Eds), 229–249. New York NY: Academic Press.

Austin, John L. 1975. *How to do things with words,* 2nd Edn. J.O. Urmson & M. Sbisà (Eds), Cambridge MA: Harvard University Press.

Bates, Elizabeth & Brian Macwhinney. 1979. A functional approach to the acquisition of grammar. In *Developmental pragmatics*, E. Ochs & B. Schieffelin (Eds), 157–211. New York NY: Academic Press.

Boldwin, Dare A. 1991. Infant's contributions to the achievement of joint reference. *Child Development* 62: 875–90.

Brooks, Patricia & Michael Tomasello. 1999. How young children constrain their argument structure constructions. *Language* 75: 720–738.

Burdelski, Matthew. 2006. Language socialization of two-year old children in Kansai, Japan: The family and beyond. Ph.D. dissertation, University of California, Los Angeles.

Clancy, Patricia. 1985. The acquisition of Japanese. In *The crosslinguistic study of language acquisition,* Vol.1, D.I. Slobin (Ed.), 373–524. Mahwah NJ: Lawrence Erlbaum Associates.

Clancy, Patricia. 1996. Referential strategies and the co-construction of argument structure in Korean acquisition. In *Studies in anaphora* [Typological Studies in Language 33], B. Fox (ed.), 33–68. Amsterdam: John Benjamins.

Clancy, Patricia. 2003. The lexicon in interaction: Developmental origins of preferred argument structure in Korean. In *Preferred argument structure:* Grammar as architecture for function, J. DuBois, L. Kumpf & J. Ashby (Eds), 81–108. Amsterdam: John Benjamins.

Clark, Eve V. & Barbara Kelly (Eds), 2006. *Constructions in acquisition*. Stanford CA: CSLI.

Fry, John. 2003. *Ellipsis and wa-marking in Japanese conversation*. New York NY: Routledge.

Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford: OUP.

Greenfield, Patricia. 1979. Informativeness, presupposition, and semantic choice in single-word utterances. In *Developmental pragmatics*, E. Ochs & B. Schieffelin (Eds), 159–166. New York NY: Academic Press.

Hashimoto, Tomoya & Shigeaki Amano. 2007. Youji-ni okeru jyoshi '*wa*'-no kakutokukatei (The process of the acquisition of the Japanese particle '*wa*'). Poster presentation at 9th annual conference of the Japanese Society for Language Sciences, Miyagi Women's Christian University, July 8.

Hatano, Etsuko. 1979. Kodomo-ni okeru jyoshi *wa*/*ga* no kakutoku no kenkyu (A study on the acquisition of the particles *wa* and *ga* by children). *Japanese journal of educational psychology* 27: 160–168.

Hinds, John, Senko Maynard & Shoichi Iwasaki (Eds), 1987. *Perspectives on topicalization: The case of Japanese wa*. Amsterdam: John Benjamins

Hirakawa, Makiko. 2004. Acquisition of case and topic particles and development index for Japanese. *Comparative research for a developmental index for first and second language of Japanese and English*, K. Otomo (Ed.), 167–174. (Report of the Grant-in-aid for scientific research (B) (1) (2001–2003) Project No. 13410034)

Kuno, Susumu. 1973. *The structure of the Japanese language*. Cambridge MA: The MIT Press.

Kuriyama, Yoko. 2001. Nihongoji-no kaku-joshi sanshutsu-no bunseki (An analysis on the productivity in the use of case particles by Japanese children). *A crosslinguistic study for the universal developmental index*, H. Shirai (Ed.), 114–123. (Report of the Grant-in-Aid for Scientific Research (A)(2)(1999–2001), No.11694009)

MacWhinney, Brian. 2000. *The CHILDES project: tools for analyzing talk,* Vol.2, *The Database,* 3rd Edn. Mahwah NJ: Lawrence Erlbaum Associates.

Miyata, Susanne. 1992. 'Papa *wa*?': kodomo-no *wa*-o fukumu shitsumon-ni tsuite ('papa *wa*?': a study on children's questions including the particle *wa*). *Bulletin of Aichishukutoku Junior colledge* 31: 151–155.

Miyata, Susanne. 2004a. *Japanese: Aki corpus*. Pittsburgh PA: TalkBank. 1–59642–055–3.

Miyata, Susanne. 2004b. *Japanese: Ryo corpus*. Pittsburgh PA: TalkBank. 1–59642–056–1.

Miyata, Susanne. 2004c. *Japanese: Tai corpus*. Pittsburgh PA: TalkBank. 1–59642–057-X.

Nagano, Masaru. 1959. Yoji-no gengohattatsu-ni tsuite: Shutoshite joshi-no shutokutatei-o chushin-ni (On language development of young children: Focusing on the acquisition process of particles). *Kotoba-no Kenkyu* (Studies in language) [The National Language Research Institute Collection 1]. Tokyo: The National Language Research Institute.

Nakamura, Kei. 1993. Referential structure in Japanese children's narratives: The acquistion of *wa*. In *Japanese Korean linguistics* 3, Soonja Choi (Ed.), 84–99. Stanford CA: CSLI.
Nakayama, Toshihide & Kumiko Ichihashi-Nakayama. 2000. What about *wa*?: Japanese *wa* in child-adult interaction. Paper presented at Second International Conference on Practical Linguistics of Japanese, San Francisco State University. April 1.
Noda, Hisashi. 1996. *Wa-to-ga* (Wa and ga). Tokyo: Kurosio.
Ochs, Elinor & Bambi Schieffelin (Eds), 1979. *Developmental Pragmatics*. New York NY: Academic Press.
Ochs, Elinor, Bambi Schieffelin & Martha L. Platt. 1979. Propositions across utterances and speakers. In *Developmental pragmatics*, E. Ochs & Bambi Schieffelin (Eds), 251–268. New York NY: Academic Press.
Okubo, Ai. 1967. *Yojigengo-no hattatsu* (The development of infants' language). Tokyo: Tokyodo publishers.
Oshima-Takane, Yuriko, Brian MacWhinney, Hidetoshi Shirai, Susanne Miyata & Norio Naka (Eds), 1998. *CHILDES for Japanese,* 2nd Edn. Nagoya: Chukyo University.
Peters, Anne. 1983. *The units of language acquisition*. Cambridge: CUP.
Sakahara, Shigeru. 1996. Roles and identificational copular sentences. In *Spaces, worlds and grammar*, G. Fauconniers & E. Sweetser (Eds), Chicago IL: The University of Chicago Press.
Scollon, Ronald. 1976. *Conversations with a one year old: A case study of the developmental foundation of syntax*. Honolulu HI: University of Hawaii Press.
Shibatani, Masayoshi. 1990. *The languages of Japan*. Cambridge: CUP.
Snow, Catharine. 1999. Social perspectives on the emergence of language. In *The emergence of language*, B. MacWhinney (Ed.), 257–276. Mahwah NJ: Lawrence Erlbaum Associates.
Takagi, Tomoyo. 2001. Sequence management in Japanese child-adult interactions. Ph.D. dissertation, University of California, Santa Barbara.
Tahara, Shunji & Takehiko Ito. 1985. Joshi *wa*-to *ga*-no danwakinoo-no hattatsu (The development of the discourse functions of *wa* and *ga*). *Japanese journal of psychology* 56: 208–214.
Tanaka, Hiroko. 1999. *Turn-taking in Japanese: A study in grammar and interaction*. Amsterdam: John Benjamins.
Tomasello, Michael. 1999. *The cultural origins of human cognition*. Cambridge MA: Harvard University Press.
Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge MA: Harvard University Press.
Tomasello, Michael & Patricia Brooks. 1998. Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics* 9: 379–395.
Yokoyama, Masayuki. 1997. *Bunpo-no kakutoku* **2**: *Joshi-no kakutoku* (The acquisition of grammar 2: The acquisition of particles). *Kodomotachi-no Gengokakutoku* (Children's language acquisition), H. Kobayashi & M. Sasaki (Eds), 132–151. Tokyo: Taishukan publishers.
Wray, Allison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language and communication* 20(1): 1–28.

Formulaic expressions in intermediate EFL writing assessment

Aaron Ohlrogge
English Language Institute,
University of Michigan

Abstract

Many works on formulaic sequences (FS) in L2 have focused on low-proficiency young learners (e.g., Myles et al. 1998) or advanced proficiency adults (e.g., Nattinger & DeCarrico 1992). Most such studies have focused on oral production. Most that have investigated FS usage in nonnative speaker writing have considered advanced learners only (e.g., Granger 1998). Thus far, there has been little work with regard to the written production of younger and intermediate proficiency writers.

This study examines 170 compositions written for an EFL proficiency test. 8 types of FS were identified and frequencies of each are compared to composition score. Correlations between these suggest several different relationships between FS use and proficiency level. Implications for assessment, instruction, and further research are discussed.

1. Introduction

Formulaic sequences play a critical role in successful communication and are extremely common in oral and written language. As countless authors have noted, such sequences often break the "rules" of language, whether through lexical abnormalities (e.g., *kith and kin*), grammatical abnormalities (e.g., *by and large*), or by an idiomatic or otherwise metaphorical meaning (e.g., *on the other hand*). As such, they present a special challenge to the second language (L2) learner, who must not only learn what form(s) and meaning(s) can and cannot be associated with a given sequence, but also how to incorporate learned

sequences into larger pieces of spoken and written discourse. In general, most interest in nonnative speaker (NNS) use of formulaic sequences has focused on oral production. This has included language acquisition by both young children (e.g., Girard & Sionis 2004; Myles, Hooper & Mitchell 1998) as well as adults (e.g., Pawley & Syder 1983; Nattinger & DeCarrico 1992).

Still, there have been some investigations into the use of formulaic sequences in L2 writing as well. For example, Yorio (1989) analyzed compositions written by ESL students and determined that although learners were aware of and made "extensive use" of formulaic sequences, they also made many lexical, grammatical, semantic, and pragmatic errors in such constructions. Yorio speculated that there was likely some direct and positive association between students' overall language proficiency and their ability to use formulaic sequences correctly. However, because he did not include specific data on the relative proficiency of his subjects, this intuition could not be substantiated in any precise way.

In another study of NNS writing, Granger (1998) compared corpora of essays written by both native and nonnative students. In her analysis, she found that while native speakers used a broad range of formulas, learners tended to greatly overuse a smaller set of specific formulas that they felt most comfortable with. Like Yorio (1989), Granger did not explore the factor of relative proficiency level among her students, so it is again impossible to know whether proficiency level could account for any variation in formulaic language use among the compositions she analyzed, if indeed any variation were present.

More recently, the field of language assessment has also begun to show an interest in formulaic language. Not surprisingly, such work has paid close attention to variations in proficiency level among language learners. Bonk (2001) described the development of a language test designed specifically to measure collocational proficiency. The test contained items based on information found in a collocation dictionary designed for language learners (Benson et al. 1997). Test items consisted of discrete sentences which required examinees to produce, in writing, one word of a two-word collocational pairing. Bonk provides extensive statistical evidence to show that these test items were valid, reliable and able to discriminate between learners of different proficiency levels. His results showed that collocational proficiency, as measured by this test, was highly correlated with other measures of proficiency, such as TOEFL scores and ESL teachers' relative rankings of the student subjects involved. Based on this work, it does appear that collocational proficiency is a valid and testable construct in second language proficiency.

At least three studies investigating qualitative differences in candidate performance in high-stakes language testing situations have considered the use of idioms and collocations by examinees. First, Hawkey & Barker (2004), reporting on a project to develop a common writing scale across multiple University of

Cambridge ESOL international certificate exams, analyzed a set of compositions written by candidates for several different exams, spanning a wide range of proficiency levels. They note, among many other linguistic features, a much higher frequency of collocations and idioms in highly rated compositions as compared to lower rated ones. They also examined the use of the pronoun *I* in conjunction with common verbs such as *think*. Although ultimately describing their findings on this issue as "inconclusive", they do provide a strongly hedged intuition, not supported by specific data, that higher-scoring candidates may use the first person more than lower-scoring candidates.

Additionally, Kennedy & Thorp (2007) examined a small corpus of compositions written for the International English Language Testing Service (IELTS) exam in order to identify specific linguistic features that characterize compositions rated at bands 8, 6 and 4. The IELTS is an international proficiency exam jointly administered by Cambridge ESOL examinations, Australia IDP, and the British Council. Candidate scores are reported in bands ranging from 1 (lowest) to 9 (highest). The authors found Band 8 compositions to have a great deal of collocational and idiomatic language present, whereas compositions rated as a 6 or 4 had far less use of such language. Similar to Granger (1998), Kennedy and Thorp also observed that lower-rated compositions heavily overused some standard transitional markers that were likely learned in composition courses. In a parallel investigation, Read & Nation (2006) analyzed transcripts of IELTS oral examinations of candidates rated at Bands 8, 6, and 4. They also found extensive use of idioms and collocations among candidates achieving Band 8 on the oral portion of the test, compared to much less idiom and collocation use at Bands 6 and 4.

2. Methodology

The following study explored what relationship(s), if any, exist between formulaic language use and L2 writing proficiency as measured by an intermediate-level language proficiency test. Specifically, the study was conducted to determine:

Question 1: What types of formulaic language are used by intermediate-level learners in a high-stakes writing examination?
Question 2: Do high-scoring and low-scoring writers use particular formulaic sequences with the same frequency?

If there are indeed differences between the types and amounts that high-scoring and low-scoring writers use, then it is likely that formulaic sequences can be a useful criterion in discriminating between different levels of proficiency – the primary purpose of most language tests.

In order to investigate these research questions, a small corpus of compositions written for a high-stakes English as a Foreign Language (EFL) exam was analyzed. The compositions were written between June 2004 and June 2005 by candidates as part of the Examination for the Certificate of Competency in English (ECCE). The ECCE is an EFL certification test published by the English Language Institute of the University of Michigan. Aimed at the B2 Level of the Common European Framework of Reference (CEFR), the ECCE is given twice a year in 25 countries located in Europe, Asia and Latin America. The University of Michigan Certificate of Competency in English can be used for personal, public, educational, and occupational purposes. The ECCE tests four language skill areas: reading (including grammar and vocabulary), listening, speaking, and writing (Testing and Certification Division, 2007).

The Writing Section of the ECCE takes 30 minutes and is comprised of a single task. Examinees are presented with a short written prompt and may respond to it in either the form of a personal letter or a personal essay. Specific instructions designed to guide the content of both the letter and essay are given as well. Candidates are free to choose either format, and both formats are assessed using the same scoring rubric.

Compositions are rated by at least two independent raters. One or more additional raters are used to resolve any disagreements between the first two raters. The ECCE scoring rubric for the writing section designates 5 possible scores, ranging from A (highest) to E (lowest). The scoring rubric designates 4 criteria upon which each candidate's writing sample is evaluated: content and development, organization and connection of ideas, linguistic range and control, and communicative effect. See Figure 1 for more details.

In order to investigate the two research questions stated above, a small corpus of 170 compositions was selected. This corpus of compositions had previously been used to develop and validate the scoring criteria for the ECCE writing section rubric depicted above. During that process, an initial set of 70 compositions was selected to represent a variety of language backgrounds and candidate scores. These compositions were then independently re-scored by a group of seven experienced composition raters employed by the University of Michigan. Raters first rated each composition individually, then met as a group to reach consensus on compositions in which score discrepancies had occurred. As part of a follow-up analysis to explore the functioning of the new scale, a second set of 100 additional compositions was later rated and negotiated in the same manner by 6 of the 7 raters who had participated in the initial study. Following this method, 6 of the compositions were rated "A", 28 were rated "B", 50 were rated "C", 54 were rated "D" and 32 were rated "E"; these scores are reasonably well distributed.

The candidates who wrote the set of compositions represent a total of 9 first language backgrounds. These are: Greek (85), Spanish (37), Portuguese (25),

SCORING CRITERIA FOR ECCE WRITING SECTION

		CONTENT AND DEVELOPMENT • Relevance of content to task • Quality of ideas used to develop the response	ORGANIZATION AND CONNECTION OF IDEAS • Arrangement of content • How language is used to link ideas	LINGUISTIC RANGE AND CONTROL • Variety and precision of grammar and vocabulary	COMMUNICATIVE EFFECT • How well communicative goals are achieved
EXCEEDS STANDARD	A	Richly develops an argument with original supporting details.	Smooth, effective arrangement and connection of ideas. A variety of cohesive devices are used effectively.	Broad range of grammar and vocabulary used accurately. If any errors are present they are minor and insignificant.	Appropriate register, awareness of audience, and establishment of context fully enhance the intended effect on the reader.
	B	Fully develops an argument with appropriate supporting details.	Appropriate and clear organization and connection of ideas. Transition markers used appropriately and not mechanically.	Good range of grammar and vocabulary; mostly accurate with only occasional errors.	Appropriate register, awareness of audience, and establishment of context help the reader to follow the text.
STANDARD	C	Adequately develops an argument. May rely on prompt for content	Ideas clearly and adequately organized. Standard connectors used appropriately but somewhat mechanically.	Sufficient range of grammar and vocabulary to fulfill the task. Errors in grammar and vocabulary do not interfere with reader's comprehension.	Adequate sense of audience and purpose for writing generally allow the reader to follow the text.
BELOW STANDARD	D	Inadequate development of argument. Content may be limited or primarily based on prompt. Some content may be irrelevant to the topic.	Simple, basic organization of ideas. Although standard connectors may be present, ideas themselves are not always connected.	A range of structures may be attempted, but grammar and vocabulary errors are frequent and interfere with reader's comprehension.	Some misunderstanding of audience and purpose and inappropriate register may have a negative effect and hinder the reader's comprehension of the text.
	E	Little or no development of argument. Content is irrelevant or taken directly from the prompt.	Minimal or no organization, connectors may be inappropriately used. Connection may not be apparent.	Grammar and vocabulary errors predominate and cause significant confusion.	Lacks audience awarweness and purpose for writing.

Figure 1. ECCE scoring rubric.

Vietnamese (10), Arabic (5), Italian (3), Romanian (2), Catalan (2), and Macedonian (1). Candidate ages ranged from 13 to 50 years of age. The mean age was 19.5 years, with a standard deviation of 7.5 years. Over half (63%) were between the ages of 13 and 18.

3. Analysis

In order to address the first research question, 70 compositions in the corpus were analyzed in order to identify multiword units of language that appeared to be formulaic in some way. Wray's (2002: 9) working definition:

> *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*

was employed as a general guideline in determining which units of language in the compositions were likely to be formulaic. While doing so, eight types of

formulaic sequences emerged. They are described as follows (all examples taken from the corpus):

1. Collocations: word pairs that occur together more often than chance or random probability would suggest. Examples include *high hopes* and *heavy fines*.
2. Idioms: Multiword sequences that have a single, standard meaning which is often metaphorical or not entirely semantically transparent. Examples include *money doesn't grow on trees* and *to cut a long story short*.
3. Phrasal Verbs: A sequence of a lexical element plus one or more particles (Crystal 2003: 352). Examples include *grow up* and *go out with*.
4. Personal Stance Markers: Expressions that signal a writer's personal view or opinion. These can be regarded as an example of what Pawley & Syder (1983) call "sentence frames". Such expressions generally occur at sentence beginnings and are common in argumentative and expository writing. Examples include i*n my opinion*, *I strongly believe* and *without a doubt*.
5. Transitions: Sequences used to signify the relationship between sections of a text. Examples include *on the one hand*, *first of all* and *in conclusion*.
6. Language copied from the prompt: A group of words that extends beyond a single noun phrase or prepositional phrase and appears identically, or with only minimal morphological variation, in both the writing prompt and the candidate's writing. G. Kennedy (2003) has noted that such borrowing is a common phenomenon among less proficient candidates, who may simply inflate their word count in an attempt to reach an arbitrary minimum essay length suggested by task instructions.
7. Generic rhetoric: A memorized expression vague and general enough to be applicable to almost any writing context. They are typically at least several clauses or sentences long; frequently they are quite complex and accurate in grammatical and/or lexical form, as the writer has had time to perfect them in advance of the actual writing assessment. On the ECCE, these may occur as standard openers or closers to a letter, e.g., *Thank you very much for taking my opinion into consideration* or *Taking all of the above into consideration, it should come as no surprise that* in an argumentative essay.
8. Irrelevant biographical information: Language used within the composition itself to state a candidate's name, age, school attended, etc. Although it could be argued that this sort of information is a useful rhetorical move in establishing a writer's identity before proceeding to a specific thesis, in reality it rarely, if ever, adds anything to a writer's argument. Furthermore, such expressions of personal identity and biography are so basic to L2 production that they offer little evidence of a learner's actual linguistic ability. Instead, they too are simply multiword units that a writer can perfect ahead of actual composition. An example from the present corpus is: *My name is Maria and I am sixteen years old.*

Next, the researcher worked with two independent coders to identify all examples of each type within the corpus. Coders were briefed about the nature of the project and given descriptions and examples of each type of sequence. In order to help familiarize them with the concept of formulaic language, coders were also provided with Wray's working definition quoted above (2002: 9). They then completed a practice set of 20 compositions not taken from the corpus. After doing so, they met with the researcher to discuss their coding. All discrepancies in identification and classification were negotiated to reach consensus. Coder One was an ESL writing instructor at the community college level who has also worked as a composition rater for the ECCE. Coder Two was a university professor who specializes in second language writing pedagogy. Both are native speakers of English.

The two independent coders each identified sequences of the eight types listed above. Each coder analyzed 85 compositions, one half of the corpus. Their results were compared with that of the researcher, who had analyzed the entire corpus. A complete set of codings was then compiled. The reliability (agreement rate) between the researcher and Coder One was .925, while the reliability between the researcher and Coder Two was .706. In order to increase reliability, only sequences which were identified by both the researcher and an independent coder were counted.

Additionally, the researcher and the two coders often differed regarding the exact difference between collocations and idioms. Many sequences identified as idioms by one coder were considered by the researcher and/or the other coder to be collocations, and vice versa. This led to a consideration of collocations and idioms as representing different points along a continuum of formulaic language, rather than discrete categories (cf. Van Lancker-Sidtis 2004). As a result of this, these two categories were combined into one during analysis. No similar discrepancies were observed between any of the other categories investigated.

4. Results

The amount and distribution of formulaic language throughout the corpus was highly variable. While some compositions contained many occurrences of certain types, others contained few or none at all. Because of this high amount of variability among the writing samples, the compositions were grouped together by grade (1–5) and the mean frequency of each sequence at each level was calculated. The relative frequency of each sequence type by grade level is shown in Figure 2. Type 5, transitions, occur much more often than any other type across all grade levels. The use of idioms and personal stance markers increases as composition grade increases, while the use of text copied from the prompt occurs much more often at lower grade levels.

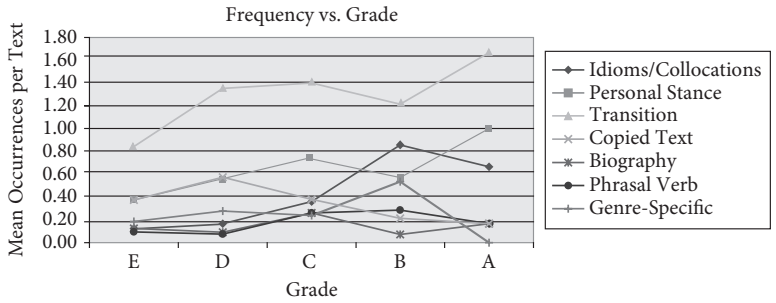


Figure 2. Frequency of FS use by grade level.

Next, the mean amounts were correlated with grade level (i.e., 1–5) using Spearman's rank correlation as a conservative and nonparametric measure. The results of this analysis are shown in Figure 3.

	Idioms & Collocations	Phrasal Verbs	Personal Stance Markers	Transitions	Copied Text	Generic rhetoric	Irrelevant Bio.
rho =	0.90*	0.60	0.90*	0.70	-0.82*	-0.10	0.10

*denotes significance at .05 level

Figure 3. Spearman's rank correlation between FS use and writing score.

Very strong and significant associations were found between grade level and two of the sequence types: idioms/collocations and personal stance markers. A correlation of .90 indicates that these types of structures occur significantly more often in compositions receiving higher grades than those receiving lower grades. The use of collocations, idioms, or personal stance markers, then, may be useful criteria for raters in determining the most appropriate level to assign a composition.

The use of transitions clearly appears to have some association with grade, given the .70 correlation between the two measure. Surprisingly, though, this association failed to reach statistical significance. This could be due to the fact that compositions rated at the B level contained not only fewer transitions than those at the A level, but also fewer than occurred at the C and D level. Further research is needed clarify why this could have occurred. Similarly, the data suggest that phrasal verbs may be somewhat more associated with compositions receiving higher scores, although it is again unclear why an association of .60 did not reach significance either.

On the other hand, the use of language copied directly from the writing prompt is quite negatively, and significantly, associated with grade. It appears that this is a rather unsuccessful composition writing strategy, or at least one used primarily by lower proficiency students. Little association in either direction is noted with regard to generic introductions and/or conclusions, and to irrelevant biographical material. It does not appear that the use of these types of sequences have much relation to composition score.

5. Discussion

As in previous studies (Kennedy & Thorp 2007; Read & Nation 2006; Hawkey & Barker 2004; Bonk 2001), the use of collocations and idioms has been shown to be strongly associated with higher proficiency students. The higher the writing proficiency grade obtained, the more likely candidates were to use these linguistic features. Additionally, personal stance markers are shown to be highly correlated with language proficiency. This particular finding should help to shed some light on Hawkey & Barker's (2004) "inconclusive" findings regarding such structures, at least as far as intermediate learners are concerned. It should be noted that Hawkey & Barker's work dealt with writing samples taken from exams at multiple levels, whereas the present study dealt with a single, intermediate level.

In the present study, there was a positive association between use of transitions and grade (rho=.70). At first glance, this might appear to contradict Kennedy and Thorp's (2007) finding that it was lower proficiency students, those rated at bands 4 and 6, who overused, and sometimes misused, basic discourse markers. However, as suggested by Taylor (2004), IELTS bands 4 and 6 are centered around the B2 level of the CEFR, the same level that passing scores on the ECCE are. This implies that candidates receiving a band 4 or 6 on the IELTS are probably of similar language proficiency as ECCE candidates. As a result, it appears that both Kennedy and Thorp and the present study have documented a similar phenomenon. Students at this intermediate level of proficiency are using (or perhaps overusing) standard transitional markers in a similar way.

Finally, while the findings presented here also support an initial expectation regarding language copied from the prompt, similar hypotheses concerning generic rhetoric and miscellaneous biographical information were not borne out. The data suggest that these phenomena are not strongly associated in either direction with overall written language proficiency, at least not at the intermediate level. These findings are supported by Wray (personal communication, April 19, 2007), who also notes no clear association between generalized rhetorical language and IELTS score.

It should be acknowledged that the categories used in this study are, to some degree, specific to this particular writing context, the ECCE. All were established using criteria intuited by the author based on experience as a composition rater and rater-trainer. However, the high agreement rates between the researcher and the coders indicate that such categories are valid and comprehensible to other researchers, at the very least those familiar with second language writing pedagogy and assessment.

Similarly, it should be acknowledged that the ages and first language backgrounds of the candidates involved in this study are not evenly distributed: there are 3–4 times as many Greek L1 candidates as Spanish or Portuguese L1 candidates, and more than half of the candidates are normally distributed between the ages of 13 and 18, while the rest are fairly evenly distributed between the ages of 19 and 50, a much wider range. However, there is no specific reason to believe that either L1 background or age have any particular bearing on formulaic sequence usage, so it is unlikely that either of these variables would have affected the results in any particular way.

6. Conclusion

This study was conducted in order to determine the answer to two research questions. Question 1 asked what types of formulaic language were used by intermediate level learners in a high-stakes writing examination. A total of 8 types of formulaic or memorized language emerged: collocations/idioms, phrasal verbs, personal stance markers, transitions, language copied from the prompt, generic rhetorical phrases, and irrelevant biographical information.

The second question concerns whether high and low-scoring writers use the same types of FS. The answer to this question is somewhat mixed. While some types were employed by both high and low scoring writers, others were more distributed. For idioms, collocations, transitions, phrasal verbs and personal stance markers, the results indicate that they are used more often by higher-scoring writers. For sequences of text copied from the writing prompt, the results indicate that lower-scoring writers are the primary users of this type of language. Finally, for generic rhetorical phrases and irrelevant biographical information, there was no clear relationship between their use and learner proficiency level.

The results here clearly demonstrate that higher and lower proficiency students at the intermediate level are using different formulaic sequences in different ways and at different frequencies. While some types of formulaic sequences are closely associated with higher proficiency students, others show little such association, and one is clearly associated with lower proficiency students. This type of information may be useful in the development and revisions of scoring rubrics for

writing tests, particularly at the intermediate level. Scoring rubrics are designed, as much as possible, to accurately describe the linguistic features salient at each level. Collocations, idioms, and personal stance markers, shown here to be positively associated with grade, are clear signals to raters that a higher grade may be most suitable for a given composition, while the use of copied text, shown here to be negatively associated with grade, should also be taken as a signal that a lower or failing grade may be appropriate. Of course, such information is to be taken into account holistically while considering all other factors described on any particular scoring rubric.

This information should also be of interest to test-taking populations and the instructors that prepare them. The present study fully supports the inclusion of collocations, idioms, personal stance markers, phrasal verbs, and transitional phrases in the L2 writing classroom. Students interested in improving their writing proficiency may be encouraged to incorporate more of these types of language into their writing. This should have the effect not only of broadening their linguistic range, but also improving their discourse organization skills; both of these features are explicitly included on the ECCE scoring rubric.

There are, of course, additional questions that must be addressed. Particularly, it isn't yet known just what effect, if any, the use of formulaic sequences has on composition raters' judgments. To what extent, if any, are raters consciously or unconsciously influenced by formulaic sequences in student writing? There is little doubt that composition raters are influenced by, or at the very least conscious of, language errors in student writing. However, do raters respond any differently to grammatical, lexical, or semantic errors than they do to errors in novel constructions? If so, how? Answers to these and other questions should help to highlight salient aspects to focus on in L2 writing instruction as well as further illuminate the role that formulaic sequences play in second language assessment.

Acknowledgments

The author would like to thank Nick Ellis, John Swales, Charlene Polio, Susan Berendes-Wood, Jean Campbell and Eric Frey for their feedback and assistance with this project.

References

Benson, Morton, Evelyn Benson & Robert Ilson. 1997. *The BBI dictionary of English word combinations*. Amsterdam: John Benjamins.

Bonk, William J. 2001. Testing ESL Learners' Knowledge of Collocations. In *A focus on language test development: Expanding the language proficiency construct across a variety of tests* [Technical Report #21], T. Hudson & J.D. Brown (Eds), 113–142. Honolulu HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Crystal, David. 2003. *A dictionary of linguistics and phonetics.* Malden MA: Blackwell.

Girard, Marie & Sionis, Claude. 2004. The functions of formulaic speech in the L2 class. *Pragmatics* 14(1): 31–53.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae, *Phraseology: Theory, analysis, and applications,* A.P. Cowie (Ed.) 145–160. Oxford: Clarendon Press.

Hawkey, Roger & Fiona Barker. 2004. Developing a common scale for the assessment of writing. *Assessing Writing* 9: 122–159.

Kennedy, Chris & Dilys Thorp. 2007. A corpus-based investigation of linguistic responses to an IELTS Academic Writing task. In *Studies in Language Testing,* 19: *IELTS collected papers*, L. Taylor & P. Falvey (Eds), 316–377. Cambridge: CUP.

Kennedy, Graham. 2003. Not so fresh in the mind: A forensic linguistic analysis of suspected memorized narrative essays. *The International Journal of Speech, Language and the Law* 10(1): 75–101.

Myles, Florence, Janet Hooper & Rosamond Mitchell. 1998. Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning* 48(3): 323–363.

Nattinger, James R. & Jeanette S. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: OUP.

Pawley, Andrew & Francis H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication,* J. Richards & R. Schmidt (Eds), 191–225. London: Longman.

Read, John & Paul Nation. 2006. An investigation of the lexical dimension of the IETLS speaking test IELTS Research Reports 6, P. McGovern & S. Walsh Canberra (Eds), *IELTS Australia* (2006): 207–231.

Testing and Certification Division, English Language Institute, University of Michigan, Ann Arbor MI, U.S.A. 2007 *ECCE Information Bulletin.*

Van Lancker Sidtis, Diana. 2004. When novel sentences spoken or heard for the first time in the history of the universe are not enough: Towards a dual-process model of language. *International Journal of Language Communication Disorders* 39(1): 1–44.

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

Yorio, Carlos A. 1989. Idiomaticity as an indicator of second language proficiency. In *Bilingualism across the lifespan: Aspects of acquisition, maturity and loss*, K. Hyltenstam & L.K. Obler (Eds), 55–72. Cambridge: CUP.

Connecting the dots to unpack the language

Ann M. Peters*
Department of Linguistics, University of Hawai'i

Abstract

Assuming each child constructs his own internal representation of the ambient language based on his experience, this process unfolds as learners take isolated linguistic chunks and discover how they relate to each other in a 4-dimensional space with phonological, semantic, pragmatic, and structural dimensions. Developmental histories of interesting early chunks reveal detours away from adult analysis and help expose the process: ***Down!***, a seemingly simple pragmatic request, must be integrated into the whole system, including its semantic opposite, ***up!***, as well as verbs it can combine with; ***please*** and ***thankyou*** are first analyzed pragmatically, with the ***you*** in ***thankyou*** being a late discovery; idiosyncratic ***whatta*** and ***didja*** must be segmented and their constituents discovered, ungrammaticizing them as they are unpacked.

1. Background and assumptions

Joan Bybee begins an article in *Language* by saying, "A usage-based view takes grammar to be the cognitive organization of one's experience with language" (2006). As an acquisitionist, this led me to the following questions:

- How do children acquire enough language to (re)organize in the first place?
- What are the starting points like?
- On what bases do they make connections?

*I wish to thank Terry Klafehn, Lise Menn, Paul Seaman and Anna Siyanova for their encouragement and comments as I developed the thoughts in this paper.

- What sorts of configurations can early sets of connections assume?
- How do connections expand?
- Are there early "islands" of connections which eventually coalesce?
- How do these (sets of) connections become more general (abstract), to the point where linguists might say "Here is a real **grammar**"?

I will present some tentative answers in this paper.

It is now well established that as they are acquiring their first language children pick up and memorize useful bits of language in a range of sizes, from adult words to unanalyzed chunks. These initial starting bits are what I think of as the "dots" that need to be organized and connected as learning progresses. When formulatic chunks are among the early dots, what ultimately happens to these them? Are they simply abandoned in favor of more analyzed language, possibly remaining as unanalyzed fossils? Or are they are mined for their grammatical information? Wong Fillmore (1976, 1979) has documented the mining process for second language learners. Pawley & Syder (1983) show that, for adults, some of the intermediate forms embedded in complex chunks are stored as partially assembled "ready-mades" or "prefabs" to be called up as needed.

I propose that this sort of multiply represented repertoire is **created developmentally** in the process of generating early connections between the initially isolated dots of language as they are encountered and learned. I begin with six working assumptions:

1. An individual's representation of language is the result of experience with ambient language use in a reasonably homogeneous community (Bybee 2006). This assumption follows in the footsteps of recent realizations that the developing brain wires much of itself in response to experience.
2. Each learner **constructs** an internal representation of language with minimal guidance from innate mechanisms (Goldberg 1995; Tomasello 2003).
3. Because they are dependent on early experience, children's first representations of language can be expected to differ quite widely from each other. But with increased exposure to the ambient language they eventually converge on the representations of the adults they talk with (Lieven 1978; Nelson 1973, 1981; Peters 1983; Pine & Lieven 1993).
4. Although an adult's representation of language has dense internal interconnections in the form of at least four overlapping networks (phonological, syntactic, pragmatic, semantic), these networks are (and must be) open (i.e., not fully connected) at the edges.
5. Language networks in the process of being constructed are far less dense than those of adults and have many more loose ends. In fact, they start out as sets of unconnected "dots" (or "islands") that are slowly connected in various ways (Braine 1976; Cruttenden 1981; Ewing 1984; Peters 1995; Tomasello 2003).

6. The less densely connected parts are loci for growth of the networks and changes in the representational system. Language grows both by adding new points and by adding new connections.

An assumption I now reject is one that I was taught in graduate school, namely that human language is best described as a completely interlocked system with every element participating in a tight set of relationships with other elements of the system. Such a network allows for no loose ends that do not participate in this tight net. Although I still think it plausible that the central core of the syntactic and phonological systems of language may be tightly interconnected in this way, I now believe that there are also elements on the periphery for which the connections are much looser, sometimes minimal. In fact, there **must** be loose ends and fuzzy edges, or how else could such tightly interconnected systems be able to grow and change? And we all know that they do: not only did Latin evolve into at least five separate languages, but new variants on the linguistic forms we are familiar with are constantly appearing. Historical linguistics tells us that some of these variants may eventually be incorporated into the grammatical system of a language. This process of grammaticization takes place when a word or phrase comes to be used in a predictable manner to signal some sort of grammatical function for which there is currently no marker (Bybee 1988; Hopper & Traugott 1993; Traugott & Heine 1991).

It seems reasonable that a child's developing language ability will also come to have a tight systematic center with loose edges for growing. The difference is that, especially at first, the proportion of linguistic forms that have been tightly linked is much smaller for the child. I have long conceptualized this process as a shift from learning initially unrelated items to their gradual interlocking into a system (i.e., from item-learning to system building; Cruttenden 1981; Peters 1986). Assuming a child begins with a set of unrelated items, how is it possible for a child to begin construct an interlocked system, much less one that resembles those of the adults around her?

The learner begins by grasping at whatever loose ends present themselves, with the driving pressures being pragmatics and phonological salience. In other words, the first language chunks to be learned are either functionally useful, or easy to hear, or both, although they are also likely to be peripheral to the more tightly interconnected (grammaticized) core of adult grammar. Ninio (1993) is the first writer I have read to ask what kinds of words are the first to be produced. She presents evidence that learners indeed begin with forms such as interjections, vocatives, and formulaic expressions, which are constrained by very few of the syntactic combinatorial properties of adult language.

Starting from items such as these, the learner builds an ever more tightly connected system fueled by the discovery of ways to connect each form along one or more of four dimensions: pragmatic, semantic, phonological, and syntactic. An important theme running through this paper is that researchers cannot really

understand how language acquisition takes place without considering how it fits in a large canvas that includes social and pragmatic development, as well as phonology, semantics and syntax. Depending on the item and the child's experience, certain similarities will be noted earlier than others. The implication is that the learner is simultaneously creating four nets, each representing similarities in one of these dimensions. These nets also become increasingly cross-connected. (This eventually accounts for adults' abilities to retrieve words in multiple ways and to solve crossword puzzles). In the process, what were once unanalyzed chunks evolve into frames with slots of ever increasing complexity and richness, and eventually into increasingly abstract representations (Feldman & Menn 2003; Abbott-Smith & Behrens 2006).

This process is not instantaneous or easy. I still believe that the tasks the learner must perform are those I proposed in *The Units of Language Acquisition* (Peters 1983):

1. **Extract** and remember chunks from ambient speech.
2. **Compare** them with other known chunks.
3. **Connect** them with chunks that seem similar in one or more ways:
   - **pragmatic**: under what circumstances is it usable? what other chunks do I know that can be used in similar circumstances?
   - **semantic**: what can I use this to express? can I connect it to other chunks that seem to have similar meanings.
   - **phonological**: what else does this sound like? how does it differ?
   - **syntax**: does it have identifiable subparts? what can it be combined with? what other chunks do I know that have similar subparts that can be similarly combined? can I make useful abstractions (beginning with limited scope formulas proposed by Braine 1976).
4. **Unpack** the chunks into recognized subparts plus any leftover bits.
5. **Store** the end products (of whatever size) in the lexicon (useful intermediate chunks remain, others fade away), some may be fuzzier than others for a while.
6. **Try out and revise.**

2. Evidence of unpacking

I will illustrate this process with evidence taken from a longitudinal database that I have been working on for a number of years (e.g., Wilson & Peters 1988; Peters 1993, 1995; Peters & Menn 1993). The child, whose name is Seth, was recorded by his father, Bob Wilson, whom I will refer to as Dad. The data base is denser than most, with from 1 to 3 hours of audio recording per week. Seth's father also kept

diary notes which track the leading edges of Seth's language productions, whereas the tapes generally reflect the middle of the road at any point. For this paper I am drawing from 40 consecutive hours of transcribed data (1;6.14 through 2;2.01), plus 30 more half-hours spaced throughout the rest of the corpus (which extends to 4;9). [Ages are shown as year;month.day.]

Seth is visually impaired, although not completely blind. This leads him to rely heavily on language to achieve social and pragmatic ends. Much of his early language production starts with the extraction of useful (and frequent) chunks from Dad's input. In the process he acquires prefabs that have considerable morphosyntactic information embedded within them. These include pronouns plus modals: *didja, do'ya, are'ya, can'ya, let's, lemme, sh'we, dontcha, wouldja, whatcha, umma*; modals plus *to*: *wanna, gonna, hafta, gotta, liketa*; and locatives or demonstratives with forms of *be*: *here's, there's, where's, this's, that's, what's, it's*. All of these adult combinations are first acquired by Seth as unsegmented units that cannot unequivocally be identified with specific adult targets.

Until about 1;10 Seth often precedes a lexical item with a schwa or a syllabic nasal, e.g., *a hot, n tape*, (Peters & Menn 1993). These bits are so indeterminate that I have called them Fillers (Peters 2001a). Fillers are not unique to Seth; they have been reported in Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Sesotho, Spanish, and Swedish (see Peters 2001a for a review). A number of researchers (e.g., Dore, Franklin, Miller & Ramer 1976; Veneziano & Sinclair 2000) propose that Fillers serve children in bridging from one-word to two-word utterances, through extension of utterances to make them sound more adult-like.

Between 1;7.10 and 1;9.01 40–50% of Seth's utterances contain Fillers (Figure 1). Then, rather suddenly, the rate of production drops rather dramatically to less than 20%, remaining low for nearly a month. We refer to this as his Filler Drop (Aoyama, Peters & Winchester, in submission). Beginning at 1;10.05 their rate increases again, and at the same time they slowly become increasingly identifiable with smaller and smaller subsets of adult morphemes. At this stage, following Dressler & Karpf (1995), we call them Protomorphemes. In addition to changes in the percentages of utterances with Fillers, the positions of Seth's Fillers change as well. (See Figure 2.) Between 1;6.26 and 1;9;01 more than 98% of his Fillers precede a single Lexeme (fL and fLL, in Figure 3, e.g., *n tree*, *n daddy*); whereas after 1;9.16 increasing numbers of Fillers begin appearing between two Lexemes (LfL, e.g., *down uh slide*). A few Fillers now also appear after verbs (fLf, e.g., *a fix'i*) and precede longer strings of Lexemes (fLLL, e.g., *a throw way cup*).

My interpretation of his Filler rebound is that by 1;10 he has amassed enough phonological and positional information about adult grammatical morphemes to be surer of what other sounds go where with respect to particular lexical items. This allows him to begin to converge on several subsets of protomorphemes, which

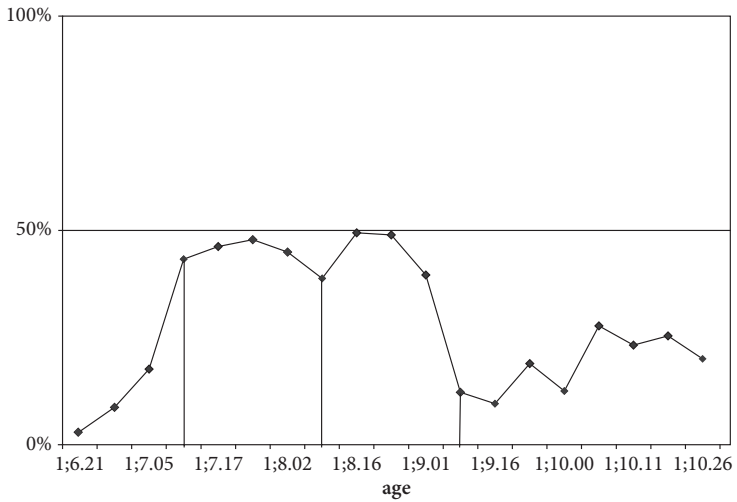


Figure 1. Percents of Seth's utterances with fillers, 1;6.21–1;10.26.

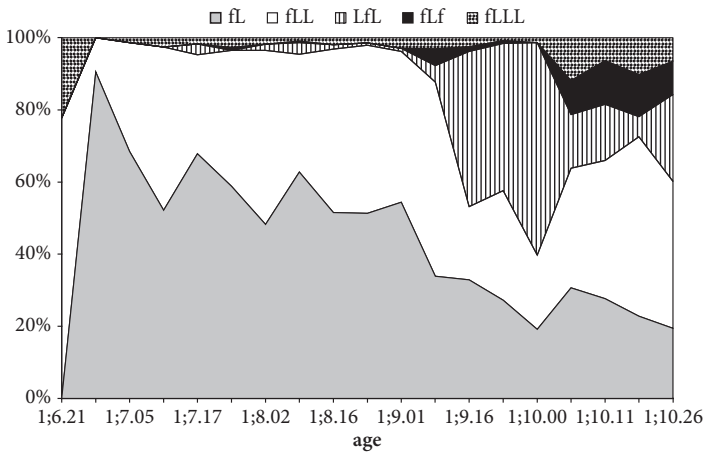


Figure 2. Percent changes in locations of Seth's Fillers, 1;6.21–1;10.26.
Note: fL = Filler precedes monosyllabic Lexeme (e.g., *m play*), fLL = Filler precedes disyllabic Lexeme (e.g., *n daddy*), LfL = Filler between two monosyllabic Lexemes (e.g., *down uh slide*), fLf = Fillers both precede and follow a Lexeme (e.g., *a fix'i*), fLLL = Filler precedes multiple Lexemes (e.g., *m play in water*).

we can consider positionally, functionally and phonologically. Positionally we can identify pre-nominal Protodeterminers (*a, the, that, this*, all approximated by *da* or *a*), and pre-verbal Protomodals (*do, did, can, let, want, gonna, should/shall*), generally fused with pronouns (*dya, aya, les, shu*). Functional classes include Question-formers (*d'ya, a(re)'ya, whatta, whatcha*), Intention-markers (*gonna, wanna, umma*), and Request-initiators (*let-Daddy, lets, shu, can'ya, can-Daddy*). One of Dad's parental concerns was that Seth express himself "politely", rather than with bare imperatives. In response to Dad's pressure, Seth seems to experiment with a range of mitigated ways of making requests, working with one form on one tape and then switching to another on a subsequent tape.

Developmentally, protomorphemes act like holding tanks for the accumulation of ever more specific information about the precise nature of their members. Ultimately Seth segments and fully acquires all the morphemes included in these amalgams, but in the process he takes some interesting detours away from the adult targets. It is these detours, where he seems to be making his own sense of the content and function of these chunks, that afford us insights into a child's process of constructing an increasingly connected language system. Let us see what we can learn from the early life histories of some of Seth's early situation-specific amalgams.

3. Forming pragmatic and semantic connections

Following Halliday (1973), I believe that the driving force that has the earliest impact on the extraction of language items is pragmatic because it enables the child to interact with adults in desired ways. Useful expressions include greetings, farewells, ways to get picked up and put down, words for handing and receiving, etc. Pragmatic and semantic connections are made by noting the range of circumstances in which a form is usable, as well as other known items that can be used in similar circumstances.

Many children associate *thankyou* with both giving and receiving because adults say it when the child gives them something. Seth starts to approximate *thankyou* about 1;8; it is a clue that he has just handed something to Dad. This continues until Seth is around 2;1 when Dad begins to teach him to say *thankyou* for favors done, using both direct instruction (*say thank you*) and modeling the appropriate usage himself.

When Seth is 1;7 Dad teaches him to respond with *please* to questions such as *do you want X?* The consequence is that for Seth *piys* becomes his way to answer *yes*. This usage starts slowly, peaks at 1;10 and then gradually drops off as Seth learns other ways to say *yes*: particularly *mhm*, *yeah* and *uh-huh*.

The ability to initiate an interaction with a caregiver is an important skill which small children are motivated to develop (Halliday 1973). On the earliest

tapes at around 1;4 Seth's dialog openers consisted of requests (to be put down *ntuu*, to be picked up *chih*, for food *babaa*, for a hug *gagaa*) and routines (*knock-knock* or animal sounds). Many interactions were initiated by Dad asking Seth questions in order to elicit language for the tape he was making. Seth also imitated a considerable number of words Dad had just said. As his vocabulary grew he developed what we have called "association bundles" of words that tended to occur together within interactive routines; e.g., *light/switch/on/push-hard*; or *tree/leaf/bark/root*; or names of relatives they had visited (Wilson & Peters 1988). Seth and Dad often engaged in interactions that consisted of rehearsing the contents of these bundles, Seth more and more often taking the lead by using a key word to initiate the exchange. Each time Seth added a word to the list Dad confirmed it.

Example (1) is from 1;8.22 when Seth led off a recitation of the names of the inhabitants of the house where his Mommy then lived: Sean, Eji, Zach, Lady and Kitty were children, roommates or animals belonging to the household. Either Mommy's own name was not part of this bundle, or he had already linked into two lists.

(1)	S:	un shan.	1;8.22	
	D:	Sean.		
	S:	reji.		
	D:	Eji?		
	S:	n za-ak.		
	D:	Zach.		
	S:	mlaydish.		
	D:	an' Lady.		
	S:	ng kitty.		
	D:	an' Kitty.		
	S:	n zah.	*starting to repeat*	
	D:	well, who else is there?	*so Dad tries to elicit more*	
		who are you forgetting.		
		how'bout your Mommy.	*suggesting another relevant name*	

We see here that even though Seth was still at the one-word stage, his control of this mode of interaction enabled him to initiate a topic and pursue it through seven interchanges. Within his limited capabilities Seth was already able to lead this conversation. (This sort of interaction is found numerous times in the next month or two.)

When Seth was 1;7 Dad took him on a six week trip to visit relatives in Texas and Oklahoma. During that period Seth's vocabulary was mushrooming, and he acquired a number of semantically linked lists: names of relatives (Tonya, David, Daniel, Ben, Julie, Erika), names for farm animals along with some associated sounds (geese/honk, chickens/chickchick, ducks, rabbits, doggie), names of foods he eats (oatmeal, orange-juice, milk), names of parts of a tree (tree, bark, roots, leaves, trunk, stump, branches), names of things in his bed (pillow, blankie, Pink

Eddie, Gonga). These lists not only enabled him to participate in dialogs such as those in (1), they seem to have provided a basis for the extraction of semantic classes. It remains to be investigated whether lists like these might also have helped him discover the grammatical class of nouns.

4. Unpacking directional semantics

Examples of the unpacking of semantic opposites are to be found in several developmental stories involving locatives because prefabs with opposing directions are not uncommon. A fairly simple story is that of *turn'on* vs. *turn'off*. At 1;7 Dad teaches Seth how to turn a light on and off, often without using any directional particle, e.g., *can you turn the switch?* Dad also talks about turning other things on and off: TV, fan, tape recorder, water. At 1;10.06 Seth gets an explicit lesson in *on* vs. *off* while he is taking a bath in the kitchen sink. His confusion of *on* and *off* is evident.

(2)	D:	you wanta stay there?	1;10.06	
	S:	piys.	*'piys' means 'yes'*	
	D:	okay.		
	S:	n take e baf?		
	D:	okay.		
	S:	n take e baf?		
	D:	okay with me-e.		
	S:	take e baf? n ta **off**? n ta **off**?		
	D:	ya'wanta turn'na water **on**?	*emphasizing 'on'*	
	S:	piys.		
	D:	okay? we're gonna turn it **on** but-		
		we're not gonna run hot water a long time		

Another fairly simple detour involves unpacking the chunks *get-it-out* and *put-it-back*. In their mealtime ritual Dad encourages Seth to help move his highchair: *let's get'chr highchair out* at the beginning, and *push your highchair back* at the end. Seth learns the first phrase, *highchair out,* but persists in using it at both ends of the meal, despite corrections by Dad.

(3)	S:	a getcha highchair out?	2;1.13
		getcha highchair out.	
	D:	okay. let's get in the highchair.	
		20 minutes pass; the meal is over	
	D:	say help me jump down..	
	S:	a jump down.	
	D:	okay. good	
	S:	ya highchair out.	
	D:	push the highchair back. push it back.	

A third semantic story concerns a hitch in the development of *down*. Like most children Seth quickly learned ways to request being put down and picked up. An idiosyncratic form *ntuu* for 'down' is already in place on the very first tape at 1;3.27, while his form *chih* for 'up' is attested at 1;4.02. At 1;7.17 he learns to say *down* and abandons *ntuu*. He gradually combines *down* with various verbs (*fall, get, jump, put*, and *sit*) and it is this last combination, *si'down*, that develops a life of its own starting at about 1;10. At 2;0 Seth begins to combine this idiosyncratic prefab with nouns: *si'down potty seat, si'down a wall, si'down a bed*. One function of such utterances is to request Dad to sit on the floor and play with him. Around 2;2 he asks Dad to join him in the bath, saying *ha(fta) si'down the shower, Daddy*, to which Dad replies *Daddy doesn't wanta get in there.* A little later, when Seth again asks *you si'down the shower?* Dad replies *Daddy doesn't wanna sidown the shower. Daddy has his clothes on.*

This father-son pair does quite a bit of roughhousing, and Seth delights in sitting on Daddy's face, often with a wet diaper. When Seth begins to call this activity *si'down face*, Dad adopts it, and the combination is used quite frequently for a short while. Is the existence of this fossilized prefab a problem? Or does it just coexist with *down*? Two small incidents at 2;1.03 shed some light on this question. Dad is sitting on his bed while Seth plays on the floor. Seth wants to join Dad on the bed, but doesn't know quite how to make the request and Dad isn't quite sure what Seth wants. The conversation goes like this:

(4) S: lie down? n si'down?
D: y' wanna-
S: n set down? *syllables distinctly spaced, then they speak together*
{n xxx?
D: {ya sit down. you wanna k- sit down on the bed.

A little later Dad is lying on his bed and Seth wants to join him. Dad realizes that Seth wants to be set up on the bed and takes the opportunity to give him a language lesson:

S: a si'down? *wants to get up on the bed with Dad again*
un si'down. *2 sec; S mutters unintelligibly; overlapped by D*
{a s- it down, Da'y.
D: {don't say sit down.
say I wanta get up.
say help me get up.
S: wanta ge'- *3.7 sec; S breathing hard; overlaps D's next utterance*
D: when you say si'down nobody knows what'chu mean.

Still later, after Dad gets up to do something, he announces that wants to sit back down on the bed. He suddenly realizes why Seth has been phrasing his request the way he did and provides another language lesson:

D: le's'go si'down on'the bed. *2.3 sec silence*
{oh that's why you say si'down on'the bed? *overlapped by S*
S: {umma be-ed.
D: 'cause Daddy says si'down on'it. *rapidly, mostly to himself*
but it's down f'r Daddy.
{it's not down f'r you. it's up, *overlapped by S*
S: {un te yao button, Daddy?
D: yeah come'on up here. come'on si'down with Daddy.
it's alright. I'll let'chu say si'down on'the bed. *3.4 sec*
see 'cause you hafta get up ta get on'the bed.
an' Daddy hasta get down. *1.1 sec*
that's why Daddy says it that way.

5. Unscrambling phonology and morphosyntax: The case of *whatta*

A relatively simple example of an unpacking "detour" is Seth's idiosyncratic *wh*-form, *whatta*. I have identified four likely sources for *whatta* in Dad's speech, all of which occur regularly across the tapes:

(5) *what are* as in What're {you/we} {doing/gonna do}.
what do as in Wha'da {you want/we have here}.
what did as in Wha'did {ju/we} {eat/say/do}.
what a as in What a {nice/funny} {kiss/song/present}.

The problem to be overcome here is that all of these forms not only sound very similar but also serve somewhat similar discourse functions. Thus it is not surprising that a learner might provisionally conflate them by linking them phonologically. Moreover, three of these forms (excepting *what'a*) occur regularly with *we* and *you*. They not only must have posed a segmentation problem for Seth, but also have contributed to his problems in sorting out the ***be*** and ***do*** auxiliaries. Up until about 2;8, his productions sound unremarkably like informal adult speech with contracted auxiliaries:

(6)

Seth	**target**	**age**
whatta you drinking.	what are	2;0.00
whatta Mommy say.	what did	2;3.11
whatta we do.	what did	2;6.00
whatta we eat.	what do	2;6.20
whatta you gon buy.	what are	2;8.06

However, between 2;8 and 3;2 he uses *whatta* in constructions that sound anomalous to adult ears:

(7) a. *whatl we're gonna buy at the store.* 2;8.06
b. *what'r we gon' buy.*
c. *whatl we gonna buy.*

d.	*what'r you gon' buy.*	
e.	*whatta you gon' smell first.*	
f.	*whatta you're gonna smell.*	
g.	*whatta we saw at the zoo.*	
h.	*whatta we're going.*	
i.	*whatta 'light' means.*	2;9.06
j.	*whatta we do at Kailua beach.* (twice)	2;10.00
k.	*whatta it has on da ginger book.*	2;11.02
l.	*whatta I was doing on my plastic bag.*	
m.	*whatta I were doing there.*	
n.	*whatta I wearing.*	3;2.09
o.	*whatta are we wearing?*	
p.	*whatta you put your head on.*	
q.	*whatta I'm lying on.*	

Lines (g, i, j, k, n, p) sound like they are missing an auxiliary; lines (a, c, f, h, l, o, q) sound like they contain two, while line (m) sounds like it contains the wrong one. Minimal pairs of utterances (a/b, e/f, l/m, n/o) reinforce the impression that he is not sure whether an auxiliary is needed or not, and if so, which one. What is going on? After careful listening to Dad's pronunciation, I have concluded that Seth extracted and identified his own idiosyncratic *wh*-form, *whatta,* which persists for several months (Peters 1993, 2001b). It is only as he identifies the auxiliaries *do*, *did* and *are* as separate morphemes with unique phonologies and distributional properties that he sorts all this out. At that point *whatta* can lose its autonomy and recede into the arena of informal speech.

6. More phonology and morphosyntax: Unpacking *didja*

A somewhat more complex detour involves *didja*. This is a form that was frequent in the input because Dad developed a habit of asking Seth semi-rhetorical, "known answer", questions with two functions. The first was to comment on something ongoing or that that had just happened:

(8)	**Input**	**Seth's age:**
	didju wanta feel'it so'more?	1.6.21
	didju toot in y'r pants?	1;8.02
	didju learn'ta stand up by yourself?	
	didju jus' wake up?	
	didju take me'cine?	1;10.0
	didju throw it on the floor?	

The second function was to encourage Seth to remember and talk about something they had recently done together:

(9)	**Input**	**Seth's age:**
	didju see trees?	1;6.21
	didju talk'ta y'r Mommy on'the phone?	
	didju enjoy those balloons?	
	didju eat icecream las' night?	
	dju hear a firetruck today.	1;8.02
	didju go in the pool yesterday?	
	didju splash in the water?	
	didju take a nice bath?	1;10.0
	didju cutchr fingernails?	2: 0.00
	didju get a new phone?	

As Seth got older and did more things out of Dad's presence, Dad began using *didja* to quiz him about what he had done:

(10)	**Input**	**Seth's age:**
	didju have breakfas' with Mommy	1;11.25
	dju go ta Mommy's house?	
	wha'didju an' Mommy do.	
	wha'didju have'at school today.	2;0.01
	didju have cheese-- have hamburger in school?	
	did they show you't school howta build a house wi'the blocks?	2;0.24
	didju have diarrhea at at school today?	
	didju stay awake at Myrna's house so late las' night?	2;2.01
	when you went with Myrna an' Lyle didju get'a hamburger?	
	didja eat'it all up?	
	didja have'a good time with Myrna an' Lyle?	

What sense does Seth make of this pattern of input? First, he seems not to have perceived any question force in Dad's utterances with *didja*, but to have understood them as **statements** about his personal situation. His own early productions seem to have exactly this force, as when at 1;11.25, on just waking from a nap he says: *didl wake up*. In a similar vein, he calls Dad's attention to ongoing events with statements such as *didja hear car, didja burp, didja drop it*. When he gets old enough to report to Dad about his activities out of Dad's presence, he also uses *did* and *didja*.

(11)				
	S:	did Myrna give you watermelon.	*initiating the topic*	2;6.20
	D:	well, good. Myrna told me that you like grapes. or Nancy did, somebody did.		
	S:	*diy* eat watermelon and strawberries?		
	D:	yeah?		
		a bit later		

S: *did you* go to Mommy's new house? *initiating this topic too*
D: yeah you went to-
S: we- *diya- diya* go on the waterbed?
D: oh, is there a waterbed at Mommy's new house?
S: yeah? is it fun?
D: was it fun? didju bounce on it?
S: yeah? *diya- didja* drink water?
D: didja drink water?
S: *di'you* drink ub water a' Mommy's new house?
D: did you drink water at Mommy's new house? yeah.
S: and *didju* make shishi at Mommy's new house .
D: you made shishi at Mommy's new house.
S: and play at Mommy's new house. *drawn out, ritualistic singsong*
D: yeah.

In addition to *didja*, starting at about 2: 0, Seth began producing several versions of a pre-verbal protoauxiliary *did/do*, approximated by *da, di,* and *diya*. Of course, at the same time as Seth is working on *didja*, he is having to sort out the other *do* auxiliaries, *does* and *did*. His early productions of *da/di/diy/diya* are ambiguous as to whether they derive from *do* or *did*. The numbers of tokens of each that he produced between 2;0.0 and 2;2.10 are shown in Figure 3. In general *didja* predominates, although *did* becomes a serious contender starting at 2;1.25.

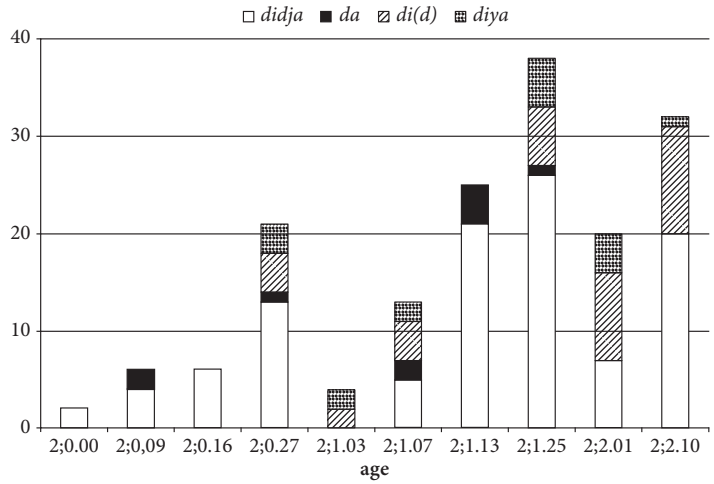


Figure 3. Tokens of *did* protoauxiliary: *didja, da, di, diya*: 1;10.05–2;2.25.

Because Bob Wilson wrote his dissertation on the development of the semantics of tense and aspect in Seth's language (Wilson 1985) I have access to Wilson's counts of Seth's use of *didja* from the diary notes. My own tape transcriptions are much more thorough than those Wilson used, but they extend only to 2;2 whereas Wilson tracked past tense development up to 3;0. To get an overview of the later development of *did*, I have used Wilson's tape counts together with his diary counts. Figure 4 suggests extreme variability and non-linearity in the development of these forms.

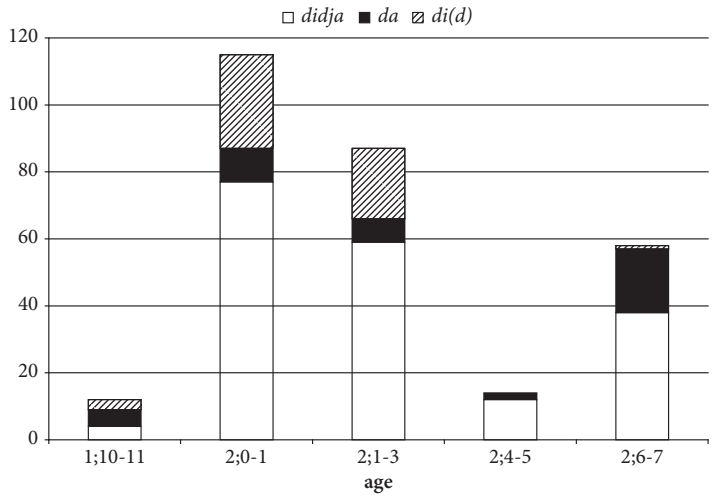


Figure 4. Tokens of *did* protoauxiliary: *didja, da, di(d)*: 1;10–2;7.

Evidence that Seth co-opted *didja* for his own purposes is his use of it in stative contexts. Here are the three examples I found in Wilson's Appendix B, along with likely interpretations:

(12)	**Seth**	**target**	**age range**
	didja have it. (the ball)	I got the ball.	26–27 months
	didja 'fraid the balloons.	I was afraid of the balloons.	28–29 months
	didja be quiet the meeting.	I was quiet at the meeting (wasn't I).	28–29 months

In adult English, neither *have* nor *be* takes the auxiliary *did*, while '*fraid* isn't even a verb! But all three utterances make sense when interpreted as **statements** about Seth's experience.

As Seth learned ever more about the phonological and combinatorial similarities and differences of the adult set of *do*-forms he was eventually able to isolate each one and integrate them all into his growing language system. (See Peters 2001b for a much fuller presentation.)

To summarize, Seth's use of *didja* as a past tense marker occurred between 1;10 and 2;10, with the auxiliary *did* completely taking over after 2;10. As with *whatta*, *didja* is then relegated to the arena of fast speech.

7. Conclusions

I started with a list of questions about how children acquire enough language to organize in the first place, and how the organizational process takes place for them. I repeat them here, together with my provisional answers.

What are the starting points like?	ITEMS OR PREFABS EXTRACTED FROM INPUT.
On what bases are connections made?	PRAGMATIC, SEMANTIC, PHONOLOGICAL, AND COMBINATORY.
What sorts of configurations can early sets of connections assume?	SIMPLE CONNECTIONS ALONG SINGLE DIMENSIONS.
How do connections expand?	BECOME DENSER AS EVER MORE ITEMS ARE LEARNED AND SIMILARITIES ARE NOTED.
Do early "islands" of connections eventually coalesce?	AS THE CONNECTIONS FOR EACH ISLAND EXPAND THEY WILL EVENTUALLY INCLUDE CONNECTIONS TO RELATED ″ISLANDS″.

The perspective I find most useful is that children extract coherent "dots" of useful language (both words and prefabs) and use them as the starting points in constructing their own language systems. The process entails the gradual unpacking and organization of dots as they are learned. The result is a set of overlapping and increasingly interconnected networks formed on the basis of similarities and contrasts in pragmatic function, semantics, sound and combinatorial possibilities.

Once researchers are willing to take the perspective that this may be what learners are doing, it is not hard to find evidence of "detours" that are unexpected from an adult perspective. Because learners must unpack prefabs that are already grammaticized, the process of recovering embedded morphemes and grammar can be seen as a kind of reverse of grammaticization. I hope I have convinced you that this is a useful perspective.

References

Abbott-Smith, Kirsten & Heike Behrens. 2006. How known constructions influence the acquisition of other constructions: The German passive and future constructions. *Cognitive Science* 30: 995–1026.

Aoyama, Katsura, Ann M. Peters & Kimberly S. Winchester. In press. Phonological changes during the transition from one-word to productive word combinations. *Journal of Child Language*.

Braine, Martin D.S. 1976. Children's first word combinations [Monographs of the Society for research in child development, 41 (1, serial No. 164)].

Bybee, Joan. 1988. Morphology as lexical organization. In *Theoretical morphology: Approaches in modern linguistics*, M. Hammond & M. Noonan (Eds), 119–141. San Diego CA: Academic Press.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82: 529–551.

Cruttenden, Alan. 1981. Item learning and system-learning. *Journal of Psycholinguistic Research* 10: 79–88.

Dore, John, Margery B. Franklin, Robert T. Miller & Andrya Ramer. 1976. Transitional phenomena in early language acquisition. *Journal of Child Language* 13: 209–218.

Dressler, Wolfgang U. & Annemarie Karpf. 1995. The theoretical relevance of pre- and protomorphology in language acquisition. In *Yearbook of morphology 1994,* G. Booij & J. van Marle (Eds), 99–122. Dordrecht: Kluwer.

Ewing, Guy. 1984. Presyntax: The development of word order in early child speech. Ph.D. dissertation, University of Toronto.

Feldman, Andrea & Lise Menn. 2003. Up close and personal: A case study of the development of three English fillers. *Journal of Child Language* 30: 735–768.

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

Halliday, Michael A.K. 1973. *Explorations in the functions of language.* London: Edward Arnold.

Hopper, Paul & Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge: CUP.

Lieven, Elena. 1978. Conversations between mothers and children: Individual differences and their possible implication for the study of language learning. In *The development of communication*, N. Waterson & C. Snow (Eds), New York NY: John Wiley and Sons.

Nelson, Katherine. 1973. Structure and strategy in learning to talk [Monographs of the Society for research in child development, 39 (1–2, serial No. 149)].

Nelson, Katherine. 1981. Individual differences in language development: Implications for development and language. *Developmental Psychology* 17: 170–87.

Ninio, Anat. 1993. Onset of speech. *First Language* 13: 291–313.

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.

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

Peters, Ann M. 1986. Early syntax. In *Language acquisition,* 2nd Edn. M. Fletcher & P. Garman (Eds), 307–325. Cambridge: CUP.

Peters, Ann M. 1993. The interdependence of social, cognitive and linguistic development: Evidence from a visually-impaired child. In *Constraints on language acquisition: Studies of atypical children*, H. Tager-Flusberg (Ed.), 195–219. Hillsdale NJ: Lawrence Erlbaum Associates.

Peters, Ann M. 1995. Strategies in the acquisition of grammatical morphemes. In *Handbook of language acquisition,* P. Fletcher & B. MacWhinney (Eds), 462–482. Oxford: Blackwell.

Peters, Ann M. 2001a. Filler syllables: What is their status in emerging grammars? *Journal of Child Language* 28(1): 229–242; 283–289.

Peters, Ann M. 2001b. From prosody to grammar in English: the differentiation of catenatives, modals, and auxiliaries from a single protomorpheme. In *Approaches to bootstrapping. Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition.* Vol. 2, J. Weissenborn & B. Höhle (Eds), 121–156. Amsterdam: John Benjamins.

Peters, Ann M. & Lise Menn. 1993. False starts and filler syllables: ways to learn grammatical morphemes. *Language* 69: 742–777.

Pine, Julian M. & Elena V.M. Lieven. 1993. Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language* 20: 551–572.

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

Traugott, Elizabeth. C. & Bernhard Heine (Eds) 1991. *Approaches to grammaticalization.* Amsterdam: John Benjamins.

Veneziano, Edy & Hermine Sinclair. 2000. The changing status of 'filler syllables' on the way to grammatical morphemes. *Journal of Child Language* 27: 461–500.

Wilson, Bob. 1985. The emergence of the semantics of tense and aspect in the language of a visually impaired child. Ph.D. diss., University of Hawai'i.

Wilson, Bob & Ann M. Peters. 1988. What are you cooking on a hot? A three-year-old blind child's 'violation' of universal constraints on constituent movement. *Language* 64: 249–273.

Wong Fillmore, Lily. 1976. The second time around: Cognitive and social strategies in second language acquisition. Ph.D. diss, Stanford University.

Wong Fillmore, Lily. 1979. Individual differences in second language acquisition. In *Individual differences in language ability and language behavior,* C.J. Fillmore, D. Kempler & W. S-Y. Wang (Eds), New York NY: Academic Press.

The effect of awareness-raising on the use of formulaic constructions

Susanne Rott
University of Illinois at Chicago

Abstract

Researchers working on formulaic language generally agree that *native-like selection* and *native-like fluency* (Pawley & Syder 1983) are attributable to the storage of formulaic constructions in the mental lexicon (e.g., Burger 2003; Wray 2002). Yet, not much is known about the incremental growth of constructions during usage events (Langacker 1987). This study assessed the facilitative effect of an awareness-raising task on the use of constructions by second semester learners of German. In addition, the study manipulated the genre of the production task (description and recipe). The study found awareness-raising positively influenced learners' use of constructions and thereby provided an opportunity for learning. Yet, the data also revealed that the level of effectiveness and the type of effect depended on the genre.

1. Introduction

Recent advances in second language acquisition (SLA) theory and a growing body of research from various related disciplines suggest that *native-like selection* and

native-like fluency (Pawley & Syder 1983) can be attributed both to the storage of formulaic constructions in the mental lexicon and to their retrieval as complete units during language use (e.g., Burger 2003; Ellis 1996; Wray 2002).[1] This knowledge has important consequences for our understanding of not only *what* needs to be learned to become a functional second language user but also *how* a second language is processed and acquired.

In order to speak (or write) fluently a language user retrieves multi-word constructions, or prefabricated lexico-grammatical units (Nattinger & DeCarrico 1992), as building blocks (e.g., Ellis 2001) to generate a message. These multi-word constructions function as "single choices" (Sinclair 1991) and are produced at a faster than normal rate of articulation. For reasons of economy, namely, to overcome the limited capacity of working memory, the use of multi-word constructions is the prevalent mode of language processing. Only occasionally does a language user switch to a word-for-word creation of an utterance (Sinclair 1991).

Many multi-word constructions have a formulaic character. That is, they have semantic and syntactic integrity, and the choice of words is often restricted (e.g., Burger 2003; Howarth 1998; Nesselhauf 2003). While some restrictions about which words can be combined are semantically motivated (e.g., *drive a car* is semantically motivated, the combination **drive a book* is not),[2] many others are arbitrary and unpredictable and therefore cannot be accounted for by grammar rules. A comparison of constructions in different languages further exemplifies the challenge for L2 learners: while in English and French (*poser une question*) one poses a question, in German one "stands" a question (*eine Frage stellen*) and in Spanish one "makes" a question (*hacer/plantear una pregunta*). Nesselhauf (2003) has pointed out that it is not sufficient for learners simply to know *which* words collocate (such as get + permission, fail + exam) but also *how* they combine (get permission **to**, fail **an** exam). Likewise, the syntactic structure of a construction representing a particular concept varies across languages. There is no reliable source for L2 learners that tells them which constructions are shared between their L1 and the L2 and are congruent and which are incongruent. The following examples illustrate the types of multi-word constructions that speakers of English, French, German and Spanish can use when they want to talk about *calling somebody*. The concept can be represented with a variety of *noun-verb-preposition* and *noun-verb* constructions, as demonstrated in Table 1.

1. Yet there is a large discrepancy how formulas as defined (see overview in Wray, 2002) and how they should be extracted from a corpus: on a purely statistical occurrences (e.g., "and the") or a combination of statistical and linguistic information (e.g., verb + noun, as in "brush teeth").

2. Yet, the idiomatic expression "that drives me crazy" has nothing to do with the act of driving a motorized vehicle.

Table 1. *Noun-verb-preposition* and *noun-verb* constructions in English, French, German and Spanish

English	to talk to X on the phone	to call X	to phone x
French	parler a X au telephone	*appeller X*	*telephoner a X*
German	*mit X am Telefon sprechen*	*X anrufen*	*mit X telefonieren*
Spanish	*hablar a X por teléfono.* *hablar con X por teléfono.*	*llamar a X*	*telefonear a X*

In order to sound *native-like* a language user needs an array of constructions from which to choose. The morphological, syntactic and semantic form of the construction depends on the particular semantic, pragmatic and discursive functions of a communicative context (e.g., Langacker 1987). The difference between informal and formal addressees, and between colloquial and formal contexts, can be expressed in the choice of words, register, and grammatical construction, as can be seen in the following examples:

a. I will call you.
b. I'll call ya.
c. I'm gonna call you.
d. I'll give you a ring.
e. I will give you a call.
f. I am going to give you a call.

Additionally, native speakers do not use just any grammatically correct construction possible, but a conventionalized version used in the speakers' community. Expressions such as *I am going to make a phone call to you* demonstrate outstanding grammatical skills; nevertheless they sound like foreignisms when compared to the conventionalized *I will call you* (Ellis 2001).

The majority of SLA studies that have assessed learners' usage of formulaic constructions have been based on advanced learners. Studies that used multiple-choice, translation, and discourse completion tasks have shown that even though advanced EAP learners have substantial knowledge about formulaic language (Adolphs & Durow 2004; Jones & Haywood 2004; Schmitt, Dörnyei, Adophs, and Durow 2004), in general, knowledge of formulas lagged behind learners' overall knowledge of vocabulary (Bahns & Eldaw 1993), their knowledge of rare words (Arnaud & Savignon 1997), and their overall language proficiency (Irujo 1993). Other studies based on free production tasks, oral interviews, and essays, have compared advanced learners' formulaic language use with native speaker data. Qualitative analyses based on a wide variety of different formulas (e.g., collocations, lexicalized language, two and three-word sequences, verb + object noun composites) have consistently shown that advanced learners lack the diversity of formulas

used by native speakers and therefore overuse the ones they know (Cobb 2003; Cowie & Howarth 1996; DeCook, Granger, Leech, and McEnery 1998; Granger 1998; Howarth 1998; Waara 2004) or the ones that serve similar functions in the learners' L1 (Granger 1998).

From the above outline it becomes clear that formulaic constructions are both indispensable and problematic for second language learners. They therefore need to play an important role in second language instruction. Even though several publications provide suggestions for instruction (e.g., Lewis 2000; Wood 2002), they lack empirical evidence concerning which type of instructional interventions foster the acquisition and use of formulaic constructions. This study was an exploratory investigation of the effect of an awareness-raising production task on the use of formulaic constructions by second-semester learners.

2. Acquiring formulaic constructions

It is generally understood that the process of connecting a formulaic construction to a meaning requires experiencing communicative events in which the construction is used. Tomasello (2000: 237) explains that "all linguistic knowledge […] derives in the first instance from the comprehension and production of specific utterances on specific occasions of use." Yet, only in a few instances does a single usage event result in the complete learning of a construction.

Naturally, multi-word constructions develop like individual words and grammatical constructions across multiple interrelated continua (VanPatten, Williams & Rott 2004) that mark partial to complete, weak to robust, and nontargetlike to targetlike knowledge. An initial encounter may result in the establishment of a memory trace of the orthographic and/or phonological form and its conceptual meaning. Subsequent processing is necessary to fill in, restructure, and strengthen the connection. For the learning of formulaic constructions, this means not only the learning of the orthographic and phonological representation of multiple words but also, in many instances, the additional encoding of morphosyntactic information. Compare the English and German construction of the concept of *to call somebody* or *to phone somebody.* In English it is represented with a *verb + [object] name* formula, or as in *to talk to somebody on the phone* which is presented with a *verb + [prepositional phrase] name + [prepositional phrase] method of communication* formula (these could also be considered two formulas: *to talk to somebody* and *on the phone*). The same concept is expressed in German (*jemanden anrufen*) with a *[direct object] noun +[separable prefix] verb* formula or with a *[dative] preposition + [object] name + verb* formula as in *mit jemandem telefonieren*.

The few SLA studies that have assessed the effect of explicit learning of formulaic constructions have confirmed that the learning process is incremental. For

example, Schmidt, Dörnyei, Adolphs, and Durow (2004) found a small but significant development over a period of two months. A qualitative item analysis revealed that for 83% of the formulas, knowledge advanced (for example, 16% from receptive to productive knowledge), yet for 17% of the items, knowledge deteriorated. Wray (2004) further found that even an explicit learning event may result in weak form-meaning connections that are subject to loss over time. In a case study in which the participant had learned a text of 60 formulaic sequences, Wray found that over the course of one week, and increasingly after five and nine months, her learner edited out particles and unstressed syllables that lacked a semantic function. In another study Fitzpatrick & Wray (2006) explored the memorization of native-like formulas that L2 learners believed to need for a future conversation. The data revealed that learners produced more targetlike formulas in a practice session as compared to a free conversation task. Interestingly, the deviations could not simply be explained in terms of an inability to memorize the formulas correctly. Instead, learners made deliberate changes in an attempt to adjust the meaning and the style of their message. They were not aware of the fixed formulaic character of the expressions.

While all of these findings are based on the learning of English as a second language, not much is known about the applicability of the results to other languages. Languages that are morphosyntactically more complex, such as German, may present different or additional challenges for L2 learners. Although there are some formulaic constructions that consist of a frame (e.g., Sinclair 1991) that allows internal lexical variation, as in *set X on fire*, *leave X behind*, *a X ago*, in English the words that combine in formulaic constructions are frequently adjacent in a sentence. This is not always the case in German. Constructions can be nested in other constructions, thereby separating components. Thus, an L2 learner of German may encounter the components of the formula *meet with friends* (*sich mit Freunden treffen*) spread over several consituents of a sentence and process them as individual words and not as a formula, such as in: *Ich treffe mich heute mit Freunden im Café* (*I meet with friends in a coffee shop today*); *Ich werde mich später mit Freunden im Café treffen* (*I will meet with friends in the coffee shop later*); *Ich habe mich gestern mit Freunden im Café getroffen* (Yesterday, I met with friends at the coffee shop). Learners who do not recognize the multi-word formulaic character of an expression may fail to fill-in and strengthen a construction that may be partially or nontargetlike encoded in the mental lexicon. Likewise, a usage event that requires the production of a nested construction may be especially taxing for the working memory of second language learners. The example above suggests that when using the present perfect and future tenses a user of German has to keep the main verb *treffen* active in working memory until all the other (intervening) ideas are produced because it is the last word in the sentence. Such linguistic complexity of formulaic constructions may require the repeated exposure to usage events in order to lead to fluent production. In fact, Bybee (2002: 112) proposes in the Linear

Fusion Hypothesis that "items that are used together fuse together." and therefore lead to stronger bonds then items that are not used adjacently in an utterance.

The research reviewed here has illustrated the linguistic complexity of formulaic constructions and some of the potential difficulties for the L2 learning process. The development of *native-like selection* and *native-like fluency* (Pawley & Syder 1983) seems to depend strongly on the repeated encounter with constructions in usage events. Yet learners in a foreign language learning setting may not have enough learning opportunities to participate in usage events to fill-in, restructure, strengthen, and retrieve constructions. Targeted instructional interventions may therefore be crucial. Focus on form activities (Doughty & Williams 1998) that direct L2 learners' attention to the individual components of a formula and its boundaries in a meaning-focused communicative event may foster the storage and retrieval of formulaic constructions. The purpose of the current investigation was to explore the effect of a task that raised participants' awareness of the multi-word formulaic character of language on the use of formulaic constructions by second-semester learners of German.

3. Research questions

The following research questions were investigated:

1. Does a prewriting task that prompts the use of multi-word constructions lead to the use of more constructions than a prewriting task that merely focuses on the production of ideas?
2. Does a prewriting task that prompts the use of multi-word constructions lead to the use of more semantically and grammatically correct production of constructions in a free writing task?
3. Does the topic/genre of the writing assignment affect the use of formulaic constructions?
4. Is there a relationship between the length of the text and the number of constructions used?

4. Methodology

4.1 Participants

Second-semester learners of German at a large Midwestern university participated in this investigation. Participants were asked for their permission to use the essays they wrote as a regular course assignment for the current investigation. Even though

all participants allowed their essays to be used, they did not submit all of their essays nor did they all complete the pre-writing task. Therefore, the data collection took place over two semesters in order to have sufficiently large cell sizes. The current data analysis was based on 110 essays: Topic 1: Living (treatment group = 27; control group = 28) and Topic 2: Recipe (treatment group = 26; control group = 29).

The basic German language sequence follows a communicatively-oriented curriculum. For each topic covered in class learners participate in several vocabulary and grammar-focused activities that are contextualized and meaning focused. The general progression of activities proceeds from input-based, to output, to free production activities.

4.2 Materials

4.2.1 *Writing tasks*

As part of the course requirements students have to submit one essay for each topic covered in the second-semester German language course. The two writing assignments selected for this investigation were the 2nd and 4th assignment out of a total of 5 writing assignments. All writing assignments in the language curriculum follow a first draft, feedback, rewrite pattern. For the current investigation only, the first draft was analyzed in order to gain insights into learners' language abilities and not their ability to respond to the instructor's feedback. The two topics chosen varied in terms of genre: The first topic on living was a description. Participants were asked to describe their dream house or apartment (Appendix A). The second topic was a recipe. Participants were asked to write down the recipe of their favorite dish (Appendix A). The description task was more open-ended in that students could use the constructions introduced in class, but the topic did not require the use of any specific constructions. By contrast, the recipe genre required the use of at least some genre-specific formulas, although it was open ended in terms of which dish students chose to describe. In addition, the recipe is a formulaically more dense text because of its structure.

For each assignment students received the writing prompt and were asked (a) to brainstorm their ideas in German (see treatment conditions below); (b) write their essay; (c) check for cohesion and coherence, and finally (d) check for grammatical accuracy. The study did not control for the time students spent on the writing task. Students completed the essay as a homework assignment.

4.2.2 *Treatment conditions*

Two treatment conditions were created which differed with respect to the brainstorm pre-writing task. In the control condition learners were simply asked to brainstorm their ideas before engaging in writing their essay. In the treatment

condition learners were prompted to brainstorm at least 10 expressions they might want to use in their essay. In addition, three examples of expressions were provided in English (Appendix A). The purpose of the construction-focused writing prompt was to scaffold the writing task. Prompting second-semester learners before writing to be aware of the formulaic character of language and the importance to produce entire constructions, as compared to single words only, was expected to foster correct use of constructions. It was assumed that a constant shift between the coherent presentation of content ideas and the activation of multi-word constructions may be too demanding for the working memory of second-semester learners and result in more incorrect use of constructions.

4.2.3 *Target constructions*

Constructions were defined as multi-word formulas with semantic and syntactic integrity. Some target constructions were congruent with English, such as in *in der Stadt* (*in the city*), while others were incongruent, such as in *auf dem Land* (*in the countryside*). Here, congruence was determined by the type of preposition used. The target constructions were multi-word formulas that had been covered in class and were useful (descriptive topic on living) and/or essential (recipe topic) to complete the writing assignments. Yet, since the assignments were free writing tasks, the participants were not required to use any specific constructions. General writing instructions in the language program emphasize that students should use the vocabulary learned in a given chapter and not look up new words. The list of possible constructions for each topic is presented in Appendix B. Constructions ranged from two to five words. While most of the constructions for the topic on living were three-word units consisting of *verb-noun-preposition* formulas, a majority of the constructions for the recipe were four- and five-word units consisting of *verb-article-noun-preposition* formulas. Considering that the length of a formulaic construction may affect acquisition (Ellis & Beaton 1996), the retrieval and use of longer constructions for the recipe topic may have presented an additional challenge to participants.

4.3 Analysis and scoring

In order to determine the number of constructions used in the essay, the list provided in Appendix B was used as a guideline. All multi-word units that appeared, in content and form, as one of the listed constructions received one point, even if the construction produced by the learner contained lexical or grammatical errors. In addition, most of these constructions are not fixed. That is, in most instances the noun can be replaced. For example, the nouns *Haus* (*house*) and *Fleisch* (*meat*) in *in einem grossen Haus wohnen* (*to live in a big house*) and *das Fleisch*

umdrehen (*to turn the mean over*) can be replaced by *Schloss* (*castle*) and *Wurst* (*sausage*) respectively.

The second measure tallied the correctly produced constructions. Semantically and grammatically correct constructions received one point. Any errors that rendered the construction nontargetlike received 0 points. Replacements of nouns as described above were counted as targetlike constructions. Moreover, word order errors were not counted as an error. As long as all words of the construction were present, and verb endings and case markers were correct, learners received one point.

In addition to tallying completely-produced formulas, the degree of accuracy of partially produced constructions were assessed. For this analysis Fitzpatrick and Wray's (2006: 46) completeness measure was adopted. The closeness of reproduction of the construction was calculated as follows: "*number of words produced with same form and function as in model target utterance ÷ number of words in model target utterance.* The stipulation that a word should have the 'same form and function' was in order to avoid counting words that happened to be identical in form to the target word but were not an instance of it." Scores of all partially produced constructions were tallied.

The final analysis assessed whether students who wrote longer essays would also use more constructions. For this measure all closed and open class words of each essay were tallied and divided by the number of constructions produced.

4.4 Procedure

During the beginning of the semester students in all second-semester German classes were asked to participate in the investigation. That meant that they allowed the researcher to use their essays for the investigation. During the semester students submitted their essays in an online digital dropbox where the researcher accessed the essays for the data analysis.

5. Results

Means and standard deviations of production scores of the construction measures are reported in Table 2. In order to answer Research Questions 1, 2, and 3, two one-way independent analyses of variance (ANOVA) were conducted. One ANOVA was conducted for the descriptive task, a second for the recipe task. In each ANOVA the independent variable was the type of brainstorm group (ideas = control group; constructions = treatment group) and the dependent variables were the number of constructions used, the number of correctly produced constructions, and the completeness of the constructions produced.

Table 2. Descriptive statistics of construction measures of treatment and control groups

		Measure		
Group	**N**	**Constructions used**	**Correct constructions**	**Completeness**
Description				
Construction	27	8.33 (2.08)	2.22 (1.69)	.63 (.08)
Ideas	28	4.93 (2.16)	1.89 (1.40)	.50 (.15)
Recipe				
Construction	26	11.23 (3.34)	2.00 (1.30)	.53 (.05)
Ideas	29	12.07 (2.90)	1.03 (.94)	.55 (.04)

Note. Construction = treatment group; Ideas = control group; Description = dream house writing prompt; Recipe = favorite recipe writing prompt.

Research Question 1 assessed whether prompting second-semester language learners to produce constructions during a brainstorming task would result in the use of more target constructions in the essay than the instruction to simply brainstorm their ideas. Findings were mixed. Prompting learners to brainstorm constructions lead to significantly more target construction use in the essay of the descriptive task $F(1,53) = 35.51$, $p = .00$, $r = .63$. The brainstorming prompt did not have the same effect for the recipe task $F(1,53) = .99$, n.s.

Research Question 2, which determined whether the prompt to brainstorm constructions resulted in grammatically and semantically correct use in the essay, showed that this was the case for the recipe task $F(1,53) = 10.12$, $p = .00$, $r = .39$ but not for the description task. $F(1,53) = .62$, n.s. A further analysis determined whether the brainstorming of constructions affected the degree of completeness if a construction has been produced partially. In the descriptive task learners who had brainstormed constructions before the writing task produced them more targetlike (complete) in the essay $F(1,53) = 16.77$, $p = .00$, $r = .48$ than learners who had been prompted to produce ideas for the brainstorm task. No such effect was found when students had to write a recipe $F(1,53) = 2.10$, n.s. In each instance of a significant finding the effect size (r) was high.

While analyzing students' individual texts the researcher noticed that in the recipe task many nontargetlike renditions included the wrong choice of words, which had not been obvious during the analysis of the descriptive writing task. A post hoc tally of lexical errors showed that in 76% of the partially-produced constructions, learners produced nontargetlike nouns or verbs, whereas in the recipe task, in only 3% of the constructions learners chose a nontargetlike noun or verb.

Descriptive statistics for Research Question 4 are reported in Table 3. In order to determine whether learners who wrote longer texts also produced more target constructions, two statistical analyses were conducted. First, in two independent

t-tests the effect of the brainstorming task on the length of texts produced was assessed. One t-test was performed for the descriptive task and one for the recipe task. Results showed no effect for the recipe task $t(53) = -.49$, n.s. and a significant effect for the description task $t(53) = -3.37$, $p = .01$, $r = -.42$. That is, in the recipe task brainstorming did not effect the length of the essay, but seemingly had a limiting effect on the descriptive task. Yet, because the effect size was very low the finding can only be described as a tendency. High standard deviations reported in Table 3 indicate that participants' length of texts varied in each group. In fact, the longest texts in each condition were at least twice as long as compared to the shortest texts.

Next, the relationship between the length of text produced and the number of constructions used was determined. Two separate Pearson correlations for the two different genres (description and recipe) showed that in the recipe task students who produced more words also used more target constructions $r = .89$, $p = .00$, while in the description task no significant relationship between the text length and the use of target constructions was found.

Table 3. Descriptive statistics of essays of treatment and control groups

		Measure		
Group	**n**	**Words produced**	**Minimum**	**Maximum**
Description				
Construction	27	110.00 (24.67)	69	170
Ideas	28	138.39 (36.53)	91	196
Recipe				
Construction	26	90.73 (26.47)	51	154
Ideas	29	94.17 (25.27)	59	164

Note. Construction = treatment group; Ideas = control group; Description = dream house writing prompt; Recipe = favorite recipe writing prompt.

6. Discussion and conclusion

The current investigation expanded previous research on formulaic constructions by exploring how the usage event (Tomasello 2003) of a free writing assignment provides an opportunity for the development of multi-word constructions. The study was based on the assumption that frequent encounters and opportunities to use constructions play a crucial role in filling-in, restructuring, and strengthening their representation in the mental lexicon. Because second language learners generally lack awareness of the formulaic character of language and therefore do not

attend to formulaic constructions (e.g., Arnaud & Savignon 1997) the facilitative effect of an awareness-raising task on the use of constructions was assessed. This study explored a commonly used instructional intervention, namely a brainstorm pre-writing task, by manipulating the level of awareness for the use of formulaic constructions. In the treatment condition the brainstorm prompt asked second-semester learners to write down formulaic constructions. In the control condition learners were only asked to brainstorm their ideas before engaging in the writing task without mentioning formulaic constructions.

The main finding of the current investigation was that an awareness-raising pre-writing task positively influenced second-semester learners' use of formulaic constructions. A prompt to brainstorm constructions before engaging in writing an essay enhanced the production of formulaic constructions and thereby provided an opportunity for learning. Yet, the data analysis further showed that the level of effectiveness and the type of effect depended on the genre of the essay assignment.

Current findings suggest that if the use of constructions is inherent to the genre, such as a recipe, language learners naturally attempt to use constructions. Raising learners' awareness for formulaic constructions in a recipe writing task did not lead to the use of more constructions. In contrast, a descriptive writing assignment for which the targeted constructions were useful but not essential (Loschky & Bley-Vroman 1993) lead to the use of almost twice as many constructions when learners' attention was raised during the pre-writing task. In other words, while the awareness-raising task *pushed* (Swain 1998) learners to use constructions, the lack of the push resulted in about 40% fewer learning opportunities of target constructions. Obviously, students in the control condition may have used other constructions than the ones targeted. However, it needs to be kept in mind that second-semester learners' abilities are very limited and a lack of awareness of the formulaicity of language may encourage the use of word-for-word L1 to L2 translations. Future research needs to further determine whether learners have the linguistic means to express themselves native-like in a free writing assignment. Such research will provide additional insights into whether instructional materials address learners' needs.

Overall, learners were able to produce only a small number the target constructions completely and native-like (between 1.03 and 2.22) when considering the length of the texts (51–196 words). Nevertheless, it was a sizable percentage of the produced constructions ranging from 8% (recipe task, control condition) to 38% (descriptive task, control condition). While the awareness-raising pre-writing task significantly contributed to producing correct constructions in the recipe task, it did not substantially contribute to more correct constructions in the descriptive task. One explanation may be that the constructions used in a recipe do not have

to undergo as many morphological manipulations to be used correctly in a sentence as constructions used in a descriptive task. For example, the German recipe structure requires the impersonal use of "one" (man) plus a verb with a third person ending.[3] The brainstorming in the recipe task may have allowed learners to focus on the form of the constructions which were then ready to use in the essay.

Even though learners were able to produce a sizable number of constructions correctly, the majority of constructions they attempted to use were nontargetlike. This finding can be interpreted in two ways. Either the target constructions were only partially encoded in the mental lexicon or learners had not established an abstract schema of the type of construction and were not able to morphologically manipulate it. In either case learners needed additional learning encounters to fill-in or restructure their current version of the constructions. These findings provide further evidence that the learning of formulaic constructions progresses along different continua: partial to complete, weak to robust, nontargetlike to targetlike. Second-semester learners had encoded some chunks of constructions but not others during previous classroom activities. Future qualitative analyses may reveal which aspects learners attend to when they encounter and process constructions. Ellis (e.g., 2006) suggests that saliency, length, and orthographic and phonological regularity may affect the encoding of chunks.

Additionally, the current study did not provide conclusive evidence for the effect of an awareness-raising production task on the filling-in and restructuring of constructions. Previous investigations have shown that additional preparation time (Foster 2001) and a more controlled production task (Fitzpatrick & Wray 2006) lead to the use of more complete constructions in oral interaction. These findings imply that instructional interventions provided opportunities for development, such as filling-in and restructuring of constructions. Similar observations were made in the current investigation. When learners had time to activate constructions before the descriptive assignment their constructions were more complete, yet not completely targetlike. Time and a focus on constructions did not however improve partial constructions when learners engaged in writing a recipe. A closer look at which type of error rendered the constructions nontargetlike revealed that, in the descriptive task, the majority of errors could be categorized as grammatical, while for the recipe task a majority of nontargetlike constructions resulted from the wrong choice of word.[4] That is, an awareness-raising

3. Current textbooks teach the use of "man" plus third person verb endings. Traditionally recipes required the use of "man" plus first person verb ending.

4. All constructions that contained a lexical error were counted. This does not mean that a construction that was counted to have lexical error did not contain a grammatical error.

brainstorm task can only lead to the retrieval of linguistic knowledge that is represented in some way in the interlanguage system. The data therefore suggests that second-semester learners had some abstract schemas of constructions to draw from. Nevertheless, they were vague and incomplete and led to the use of improved but not targetlike constructions. In contrast, the lack of lexical items may have resulted in the nontargetlike combination of words that learners had available in their mental lexicon. Again, this may have been the nature of the recipe task, which requires an array of food item specific constructions depending on the complexity of the recipe.

These findings may be further supported by looking at the length of texts that learners produced for each writing assignment. The brainstorming task seemingly had no effect on text length for the recipe genre. A particular recipe simply required a certain number of steps, whether they had been brainstormed or not. Consequently, a longer recipe resulted in a higher potential that one of the target constructions would be used. The length of text was therefore significantly related to the number of constructions used. In contrast, the awareness-raising task led to significantly shorter descriptions of a dream house. This finding may be interpreted in two ways: either awareness-raising had a limiting effect in that learners realized which ideas they were not able to produce in German because they did not have the necessary vocabulary they needed; or the brainstorming of constructions took longer and learners had allotted a certain amount of time to complete their homework.

7. Limitations

Since the current investigation was of exploratory nature, it has a number of limitations that need to be addressed in future investigations. First, this study did not control for time. The extremely high standard deviations in the length of text produced shows that some learners took more time and effort to complete the writing assignment. A timed essay writing assignment may qualify the current findings. Second, this study would also benefit from providing baseline data. An additional L1 to L2 translation task for all constructions presented in class would provide further insights into the production abilities during a controlled as compared to a free writing task. Third, this study assumed that constructions are stored as units and retrieved together during use. Yet, this study did not assess retrieval procedures. Therefore, claims are tentative and require additional research methodologies to develop further insights into mental processes. Finally, qualitative analyses, such as stimulated recall protocols, may be useful to determine the source of nontargetlike constructions.

Appendices

Appendix A

Writing assignments

Topic 1: Beschreiben Sie Ihr Traumhaus oder Ihre Traumwohnung. In welchem Bundesstaat oder Stadt mA chten Sie wohnen? Warum? (Describe your dream house or apartment. In which state or city would you like to live? Why?)
Control condition: Brainstorm your ideas before writing the Aufsatz.
Treatment condition: Brainstorm at least 10 expressions you need to write the Aufsatz: For example: live on the countryside, close to the university, live in a high-rise.

Topic 2: Mein Lieblingsgericht. Schreiben Sie das Rezept für Ihr Lieblingsgericht (My favourite dish. Write the recipe for your favorite dish)
Control condition: Brainstorm your ideas before writing the Aufsatz.
Treatment condition: Brainstorm at least 10 expressions you need to write the Aufsatz: For example: Cut cheese into small pieces, add flower, brown the meat

Appendix B

Sample target constructions
Topic1: Traumhaus (dream house)
In + type of place

Examples:

In einem grossen Haus wohnen	live in a big house
In einem Wolkenkratzer	In a skyscraper
Auf einem Boot	On a boat

Location I
Examples:

In der Stadt leben	live in the city
Auf dem Land	In the countryside
In einem Vorort	In a suburb
In Deutschland	In Germany
In den USA	In the United States

Location II
Examples:

In der Nähe von	Near the
Gegenüber von	Across from
Um die Ecke von	Around the corner from

(*Continued*)

Appendix B (Continued)

Living space	
Examples:	
In der Küche	In the kitchen
Im Wohnzimmer	In the living room
Vor dem Haus	In front of the house
Details	
Examples:	
Mit Schwimmbad	With a pool
Mit einem Garten	With a yard
Mit vielen Fenstern	With many windows
Topic 2: Lieblingsrezept (favorite recipe)	
Target Constructions	
Actions	
Examples:	
Auf den Teller legen	Put on a plate
In die Schüssel geben	Put into a bowel
Öl in der Pfanne erhitzen	Heat oil in the pan
Eier in eine Schüssel schlagen	Break eggs into a bowl
Mehl dazugeben	Add flower
Das Wasser abgiessen	Pour out the water/strain the water
In Scheiben schneiden	Cut in slices/slice
In kleine Stücke schneiden	Cut into small pieces
Auf das Brot legen	Put on bread
Auf das Fleisch giessen	Pour on the meat
Das Fleisch umdrehen	Turn the meat over
Das Fleisch anbräunen	Brown the meat
Die Soβe darübergiessen	Pour the sauce over X
Karotten hinzugeben	Add carrots
Using kitchen tools	
Examples:	
Mit dem Messer schneiden	Cut with a knife
Mit der Gabel umrühren	Stir with a fork
Im Ofen backen	Bake in the oven
In der Pfanne braten	Fry in the pan
Food	
Examples:	
Eine Scheibe Brot	One slice of bread
Ein Stück Käse	One piece of cheese
Zwei Stückchen Fleisch	Two pieces of meat

References

Adoplphs, Svenja & Valerie Durow. 2004. Social cultural integration and the development of formulaic sequences. In *Formulaic sequences. Acquistion processing and use* [Language Learning & Language Teaching 9], N. Schmitt (Ed.), 107–126. Amsterdam: John Benjamins.

Arnaud, Piere J.L. & Sandra Savignon. 1997. Rare words, complex lexical units and the advanced learner. In *Second language vocabulary acquisition*, J. Coady & T. Huckin (Eds), 157–173. Cambridge: CUP.

Bahns, Jens & Moira Eldaw. 1993. Should we teach EFL students collocations? *System* 21: 553–571.

Burger, Harald. 2003. Phraseologie. *Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.

Bybee, Joan L. 2002. Sequentially on the basis of constituent structure. In *The evolution of language out of pre-language*, T. Givon (Ed.), 109–134. Amsterdam: John Benjamins.

Cobb, Tom 2003. Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review* 59: 393–423.

Cowie, Anthony P. & Peter Howarth. 1996. Phraseological competence and written proficiency. In *Language and education*, G.M. Blue & R Mitchell (Eds), 80–93. Clevedon: Multilingual Matters.

De Cock, Sylvie, Sylviane Granger, Geoffrey Leech & Tony McEnery. 1998. An automated approach to the phrasicon of EFL learners. In *Learner English on computer*, Sylviane Granger (Ed.) 67–79. New York NY: Longman.

Doughty, Catherine & Jessica Williams (Eds), 1998. *Focus on form in classroom second language acquisition*. Cambridge: CUP.

Ellis, Nick. 1996. Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18: 91–216.

Ellis, Nick. 2001. Memory for language. In *Cognition and second language instruction*, P. Robinson (Ed.), 33–68. Cambridge: CUP.

Ellis, Nick. 2006. The associative –cognitive CREED. In *Theories in second language acquisition. An introduction,* B. VanPatten & J.Williams (Eds), 77–96. Mahwah NJ: Lawrence Erlbaum Associates.

Ellis, Nick C. & Alan Beaton, 1993. Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning* 43: 559–617.

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

Foster, Pauline. 2001. Rules and routines: A consideration if 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), 75–93. Harlow: Longman.

Granger, Sylviane. 1998. Prefabricated patterns in advanced ESL writing: Collocations and lexical phrases. In *Phraseology: Theory analysis and applications*, A.P. Cowie (Ed.), 145–160. Oxford: Clarendon Press.

Howarth, Peter. 1998. Phraseology in English academic writing: Some implications for language learning and dictionary making. In *Phraseology: Theory analysis and applications,* A.P. Cowie (Ed.), 161–188. Oxford: Clarendon Press.

Irujo, Suzanne. 1993. Steering clear: Avoidance in the production of idioms. *International Review of Applied Linguistics and Language Teaching* 31: 205–219.

Jones, Martha A. & Sandra Haywood. 2004. Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In *Formulaic sequences. Acquisition, processing and use* [Language Learning & Language Teaching 9], N. Schmitt (Ed.), 269–300. Amsterdam: John Benjamins.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar,* Vol.1: *Theoretical prerequisites*. Stanford CA: Stanford University Press.

Lewis, Michael. 2000. *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.

Loschkey, Lester & Robert Bley-Vroman.1993. Grammar and task-based methodology. In *Tasks and language learning*, G. Crooks & S. Gass (Eds), 123–167. Clevedon: Multilingual matters.

Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24: 223–242.

Nattinger, James R. & Jeanette S. DeCarrico. 1992. *Lexical phrases in language teaching*. Oxford: OUP.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Native like selection and native like fluency. In *Language and communication*, J. Richards & R. Schmidt (Eds), 191–226. London: Longman.

Schmitt, Norbert, Zoltan Dörnyei, Svenja Adolphs & Valerie Durow. 2004. Knowledge and acquisition of formulaic sequences: A longitudinal study. In *Formulaic sequences. Acquisition, processing and use* [Language Learning & Language Teaching 9], N. Schmitt (Ed.), 55–86. Amsterdam: John Benjamins.

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

Swain, M. 1998. Focus on form through conscious reflection. In *Focus on form in classroom second language acquisition*, C. Doughty & J. Williams (Eds), 64–82. Cambridge: CUP.

Tomasello, M. 2000. Do young children have adult syntactic competence? *Cognition: International Journal of Cognitive Psychology* 74: 209–253.

VanPatten, Bill, Jessica Williams, Susanne Rott & Mark Overstreet. 2004. *Form-meaning connections in second language acquisition*. Mahwah NJ: Lawrence Erlbaum Associates.

Waara, Renee. 2004. Construal, convention, and constructions in L2 speech. In *Cognitive linguistics, second language acquisition, and foreign language teaching*, M. Achard & S. Niemeier (Eds), 51–76. Berlin: de Gruyter.

Wood, David. 2002. Formulaic language in acquisition and production: Implications for teaching. *TESL Canada Journal* 20: 1–15.

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

Wray, Alison. 2004. Formulaic language learning on television. *Formulaic sequences. Acquisition, processing and use* [Language Learning & Language Teaching 9], N. Schmitt (Ed.), 249–268. Amsterdam: John Benjamins.

Can L2 learners productively use Japanese tense-aspect markers?*

A usage-based approach

Natsue Sugaya & Yasuhiro Shirai
Niigata Sangyo University/University of Pittsburgh

Abstract

This study, using acceptability judgment tests from 61 learners of Japanese, examined the productivity of learners' use of tense-aspect marking in second language acquisition.

*This paper is based on research conducted as part of Sugaya's doctoral dissertation submitted to Ochanomizu University (Sugaya 2005), except for the frequency analysis of the native corpus. We thank Kevin Gregg for his comments on an earlier version of this paper and Nick Ellis for his comments on our presentation at the UWM Formulaic Language Conference. We also thank the participants in the study for their cooperation, and those who kindly helped us recruit them. This research is partially supported by grants from the Japan Society for the Promotion of Science (Grant No. 13410034; PI-Kiyoshi Otomo, Grant No. 19720127; PI-Natsue Sugaya).

To verify whether learners productively use the tense-aspect markers, we compared the score of the individual verbs in three target contexts (simple nonpast *-ru*, simple past *-ta*, and nonpast imperfective *-te i-ru*). The results indicated that higher proficiency learners correctly chose tense-aspect forms, whereas lower proficiency learners showed verb-specific preferences. The preference was more strongly observed with verbs denoting resultative state than verbs denoting progressive meaning. The results suggest that learners gradually attain productive control of tense-aspect forms, which is consistent with the proposed developmental sequence of formula > low-scope pattern > construction (N. Ellis 2002).

1. Introduction: Rule learning vs. Item learning

The acquisition of tense and aspect has been extensively investigated in first language (L1) and second language (L2) research (Bardovi-Harlig 1999; Li & Shirai 2000; Weist 2002). It has been observed that there is a strong relationship between inherent lexical aspect of verbs and the acquisition of tense-aspect morphology, a relationship summarized in the Aspect Hypothesis (Andersen & Shirai 1994, 1996; Bardovi-Harlig 1999; Robison 1995; Shirai 1991). The Aspect Hypothesis predicts that at the early stages of acquisition, learners predominantly use past tense and perfective aspect markers with punctual and telic verbs, and progressive aspect markers with activity verbs.

Studies within the framework of the Aspect Hypothesis have emphasized form-meaning relationships, but have tended to overlook the form-form mapping mechanism, although its importance has been acknowledged (Sugaya & Shirai 2007; Shirai 2004). In other words, the association between verb semantics and morphological form was the focus of the research, and not much attention has been paid to simple form-form associations, in the context of what verb is inflected for what form. To help fill this gap, this study examines the productivity of learners' use of tense-aspect morphology in L2 Japanese by using an acceptability judgment task for sentences involving various tense-aspect forms. We address the following questions: (1) Do L2 learners use Japanese tense-aspect markers in verb-specific fashion? (2) Do L2 learners eventually attain productive control of Japanese tense-aspect markers? Based on the results, we argue that in acquiring Japanese tense-aspect morphology, learners show verb-specific patterns based on rote learning at the intermediate level, while they attain productive control at the advanced level, thus supporting the claim that learners go through the developmental stages of formula > low-scope pattern > construction (N. Ellis 2002).

The chapter is structured as follows: in the remainder of the introduction, previous research on rote vs. rule learning in L2 acquisition is reviewed. In Section 2, the semantics and acquisition of the target structure – the Japanese imperfective aspect marker *-te i-(ru)* – is discussed, and the research questions are presented.

Section 3 reports the method, and Section 4 discusses the results. Section 5 discusses theoretical explanations of the findings, and Section 6 concludes the paper.

It has been widely documented that both first and second language learners produce unanalyzed chunks (Brown 1973; Clark 1974; R. Ellis 1984, 1999; Hakuta 1974; Krashen & Scarcella 1978; Myles et al. 1999; Peters 1983; Tomasello 1992, 2003; Vihman 1982; Weinert 1995; Wong-Fillmore 1976; Wray 2002). Few would deny that the rote-learned forms can help fulfill communication needs when learners' language competency is insufficient; however, whether they feed into creative language is an open question.

Wong-Fillmore (1976) extensively investigated the issue using data from five Spanish-speaking children learning L2 English, and claimed that item-based learning evolves into rule learning. She reported how a child, Nora, unpacked chunks such as "I wanna play wi' dese" during the data collection period (Wong-Fillmore 1976: chapter 6). In contrast, Krashen and Scarcella (1978) argued that they are different processes. They argued that due to demands of early production, learners may use rote-learned items dissimilar to their analytic language and that these two different languages are running side by side, although with no transfer between them. Thus, they concluded that creative language develops independently and just "catches up" with formulaic language.

The debate continued into the 1980s and 1990s. Bohn (1986) further challenged the importance of item learning, especially challenging Wong-Fillmore's (1976) study. Using the data on naturalistic L2 English acquisition by four German children from the Kiel Project database (Wode 1981), he suggested that the evidence for the relationship between formulaic language and creative language was an artifact of methodological problems. First, he claimed that Wong-Fillmore overextended criteria for identifying formulaic speech such as *can* + PRONOUN + VERB PHRASE (e.g., *Can we take 'em home now? Can I read this one?*), which have only one lexical item in common in one position. In addition, he pointed out that seemingly formulaic language was found during games and that this may determine the use of specific structural types and specific lexical elements. He concluded that the role of formulaic language was limited to short-term production tactics. Weinert (1995), however, defended Wong-Fillmore's criteria, claiming that even with narrower criteria than those proposed by Bohn, formulaic language can also play a role in the Wode data.

Myles et al. (1999) also suggested that rote-learned items form the basis for subsequent creative language. Analyzing L2 French interrogatives produced by secondary school pupils in England, they argued that learners who were able to memorize formulaic language successfully were also the learners who were the earliest to attain productive control of the construction.

In the area of tense-aspect acquisition, it has been pointed out that it is important to investigate the relationship between item-based learning and rule learning

(Bardovi-Harlig 2002; Sugaya & Shirai 2007; Shirai 2004), as mentioned above. In a comprehensive review of tense-aspect acquisition in L2 English, Shirai (2004) pointed out that studies using production data, whether oral or written, did not always follow the prediction of the Aspect Hypothesis: for some studies, the effect of inherent lexical aspect was strongest not at the beginning stage, but was so at the intermediate level. Shirai attributed the deviation to the use of rote-learned forms in production data. In other words, when pushed to perform beyond their capacity, L2 learners at lower proficiency levels might haphazardly produce high-frequency forms before the actual relationship between the morphological form and its meaning is acquired. Shirai further suggested that L2 learners, with a higher memory capacity and possibly a lower analytic ability, tend to rely on rote-learned forms. However, in spite of the accepted importance of rote learning in the acquisition of tense-aspect markers, little empirical research has directly investigated the issue.

2. Inherent aspect and the Japanese tense-aspect markers

Before we review the Japanese tense-aspect system and the L2 studies on it, we briefly describe categories of inherent aspect of verbs, which are relevant to the present study. Unlike grammatical aspect, which is marked explicitly using linguistic devices such as auxiliaries or inflectional morphology, inherent aspect is defined in terms of the temporal properties of the situation to which the verb (phrase) refers.

Vendler's (1967) four categories – probably the most broadly accepted and the best known in L2 tense-aspect studies – are state, activity, accomplishment, and achievement. A state verb (e.g., *love, know*) refers to a situation that is viewed as continuing to exist unless some outside situation makes it change. An activity verb (e.g., *run, walk*) describes a dynamic and durative situation without an inherent endpoint. An accomplishment verb (e.g., *make a chair, run a mile*) describes a situation that is dynamic and durative, but has a necessary endpoint. An achievement verb (e.g., *die, drop*) refers to a dynamic and punctual situation. States and activities are atelic (i.e., without an inherent endpoint), whereas accomplishments and achievements are telic (i.e., involving a punctual point of state-change). Smith (1991) further proposed a fifth category, semelfactive, which is punctual and atelic in that it does not result in change-of-state (e.g., *jump, knock*).

The Japanese tense-aspect system has much in common with that of English. In both languages, all indicative predicates are marked for tense (past vs. nonpast). The past tense marker can be attached to any verb without any systematic restriction. The nonpast form normally refers to present state with state verbs, and to future action or habitual action with dynamic verbs.

With regard to aspect, Japanese has an imperfective aspect marker *-te i-(ru)*, which must be used in describing action in progress at the time of reference (as in (1)), which is similar to the obligatory progressive form *be –ing* in English. However, the semantic scope of the Japanese aspect marker *-te i-(ru)* is different from that of English. The major difference between the two languages concerns the combination with achievement verbs, which are punctual and telic. Japanese *-te i-(ru)* cannot denote a process leading up to the endpoint (e.g., *He is reaching the summit*), but can refer to a resultative state (as in (2)), since Japanese *-te i-(ru)* focuses on the duration of state that occurs as a result of the punctual event.

(1) *Ken-ga utat-te i-ru/-ta.*
Ken-NOM sing-ASP-NONPAST/PAST
'Ken is/was singing.'

(2) *Booru-ga oti-te i-ru/-ta.*
ball-NOM fall-ASP-NONPAST/PAST
'The ball has/had fallen (and is/was still there).'

Previous research on L2 studies of imperfective aspect marker *-te i-(ru)* mostly indicates that progressive meaning is easier than resultative state meaning – consistent with the Aspect Hypothesis (Andersen & Shirai 1994; Shirai & Andersen 1995) which predicts strong association between activity verbs and the progressive marker (Koyama 2003; Sheu 2005; Shibata 1999, 2000; Shirai & Kurono 1998; Sugaya & Shirai 2007). This is noteworthy because the results contradict the input frequency from native speakers of Japanese. Shirai (1995; see also Shirai & Kurono 1998) analyzed the utterances that a Japanese native speaker (NS) addressed to L2 learners and showed that *-te i-(ru)* was more frequently attached to achievements than activities (59% vs. 37%). Additionally, an analysis of a conversational corpus of Japanese NSs (Shirai & Nishi 2005) found that *-te i-(ru)* was used more often with achievements than activities (60% vs. 28% out of 518 tokens of *-te i-(ru)*).

How is this related to formulaic learning? Directly relevant to the issue of rote vs. rule learning is the observation that L2 learners of Japanese tend to rely on rote learning in the acquisition of tense-aspect forms (Kurono 1998; Sheu 2005; Sugaya 2003, 2005; Uozumi 1998). For example, in the follow-up interview of Sheu's (2005) cross-sectional study, a learner reported, "I memorized *ni-te i-ru* 'resemble' only in the *-te i-ru* form all along." Although the verb *ni-ru* showed 100% accuracy in Sheu's data, this high percentage is likely to be the result of rote learning, and therefore, most probably, the learner cannot inflect this verb in a different form (e.g., nonpast *ni-ru* or past *ni-ta*).

Another example is Sugaya's (2003, 2005) L1 Russian learner Alla. In her ten-month longitudinal data, some verbs (e.g., *siru* 'come to know,' *tuku* 'be attached') were produced exclusively with imperfective aspect marker *-te i-(ru)*, while other

verbs (e.g., *iku* 'go,' *tigau* 'differ') were rarely used with *-te i-(ru)*. Even though these verbs can be used in any of the four basic forms (simple nonpast *-ru*, simple past *-ta*, nonpast imperfective *-te i-ru,* and past imperfective *-te i-ta*), Alla demonstrated a very strong verb-specific preference.

In sum, previous research has suggested that L2 learners of Japanese demonstrate difficulty in productively using tense-aspect markers, and use them in a verb-specific fashion, similar to L1 English children in Tomasello (1992). However, these observations in L2 studies are based on anecdotal evidence or a case study, and there has been no study examining the true productivity of L2 learners' tense-aspect markers with a larger group of L2 learners.

Therefore, in this study, we investigate the following research questions:

1. Do L2 learners use Japanese tense-aspect markers in verb-specific fashion? If so, is there an effect of verb type (i.e., verbs denoting progressive meaning with *–te i-(ru)* vs. verbs denoting resultative state meaning with *–te i-(ru)*)?
2. Do L2 learners eventually attain productive control of Japanese tense-aspect markers?

3. Method

3.1 Participants

There were 80 participants, who lived in the Tokyo metropolitan area: 39 English NSs, 18 German, 18 Russian, 3 Ukrainian,[1] and 2 Bulgarian. They were recruited through flyers in various public places including universities, Japanese language schools, and student dormitories as well as through classified ads in free English papers and on the Internet. The reward advertised for participation was monetary compensation (¥1,000 = $9 at the time of the data collection) and free assessment of their oral proficiency based on the ACTFL Oral Proficiency Interview (OPI) with feedback on their performance. OPI was also used to screen the participants in order to confirm their basic knowledge of Japanese, as reported in detail below.

Nine learners, evaluated as novices based on the OPI, did not take the acceptability judgment test, because it was expected that novice learners would have difficulty understanding the test sentences. The judgment test was administered to 71 learners, who were rated as intermediate and advanced; however, the data from 10 learners were not retained for analysis. Four participants did not complete the task because of insufficient knowledge of Japanese phonograms,

1. The Ukrainian learners were balanced bilinguals of Russian and Ukrainian.

and six participants who reported lack of formal L2 instruction were excluded to control for the variable of classroom instruction. This resulted in a sample of 61 participants:[2] 17 German, 13 Russian, 3 Ukrainian, 2 Bulgarian, and 26 L1 English (35 males, 26 females). The age range was 19 to 53 (mean = 27.7).

Additionally, 21 NSs of Japanese (all female) provided baseline data for the judgment test. They were graduate, undergraduate, and non-degree students of various majors at a women's university in Tokyo (age range = 20 to 55, mean = 30.2 for 20 Japanese NSs; one person did not report her age).

3.2 Materials and procedure

3.2.1 *Acceptability judgment test*

The acceptability judgment test was designed to assess learners' knowledge of finite verb forms *-ru* (simple nonpast), *-ta* (simple past), *-te i-ru* (nonpast imperfective), and *-te i-ta* (past imperfective). Each item consisted of a short dialogue with the verb deleted, with four verb forms as choices to fill in the blank. The learners were instructed to circle all appropriate forms from among the four verb forms, as illustrated in the Appendix. This was to examine if the learners could appropriately judge in which context a verb form can or cannot be used. In other words, the task required learners to judge the acceptability of all four choices. To ensure that participants gave judgments for each item carefully, we incorporated eight items, which had two correct target forms each, so that the learners would believe that some items had more than one correct answer (see Table 1). Otherwise, participants might have decided that there was only one correct choice for each item.

The test involved nine verbs for each imperfective *-te i-ru* target context (progressive and resultative). There was no overlap of verbs between the two meaning contexts because resultative requires achievement verbs, and progressive requires activity, accomplishment, or semelfactive verbs (Shirai 2000), except for one verb (*otiru* 'fall'), which was used in both progressive and resultative (to be discussed below). Therefore, 17 verb types were used in total, which were also presented in simple nonpast *-ru* and simple past *-ta* contexts. (The target verbs are shown in Tables 3 and 4 in the Results section). To check whether learners have productive knowledge of the tense-aspect markers, we compared the accuracy score for three target contexts of each verb. (See Table 1,

2. Details of 61 participants' OPI ratings are as follows: 30 participants were evaluated as intermediate and 31 as advanced. They were distributed from intermediate-low to advanced-high. It was not possible to conduct an OPI for one Russian learner due to a scheduling difficulty, but this person was retained because her proficiency was judged to be good enough to perform the tasks.

Table 1. Test items

Target contexts	Correct forms	Number of items
Simple nonpast	*-ru*	16
Simple past	*-ta*	17
Nonpast imperfective (Progressive)	*-te i-ru*	9
Nonpast imperfective (Resultative state)	*-te i-ru*	9
Nonpast habitual	*-ru/-te i-ru*	4
Nonpast habitual	*-te i-ru*	2
Past habitual	*-ta/-te i-ta*	4

in which these items are highlighted. The other items were distracters, targeting habitual meanings). However, there are two exceptions. The verb *siru* 'come to know' was provided in simple past and imperfective *-te i-ru* target contexts, but lacked the *-ru* target. This is because the nonpast form *siru* 'come to know' is rarely used as a predicate of a main clause, and it was difficult to come up with a sentence that could be understood by non-native speakers. The verb *otiru* 'to fall' was presented in the both progressive *(suiteki ga oti-te i-masu*[3] 'drops of water are dripping' used as a semelfactive verb) and resultative state (*saihu ga oti-te i-masu* 'a wallet has fallen [and it is there]' used as an achievement verb) meanings to test the contribution of form vs. meaning, but was provided in one simple nonpast and one simple past target context (both as achievement verbs to describe a wallet falling/a person falling into a pond). This resulted in 18 items for the nonpast imperfective *-te i-ru* target, 16 in simple nonpast *-ru*, and 17 simple past *-ta* (see Table 1).

The judgment test was piloted with 13 NSs of Japanese, none of whom were in the control group of 21 Japanese NSs mentioned previously. The test items to which more than 20% of the 13 informants did not respond with our expected response were revised. The test was also piloted with three learners of Japanese, and we reworded difficult or unclear expressions that they pointed out. For all the items in the present test, more than 90% of the controls chose the same verb forms that we deemed correct.

The sentences were given in Japanese orthography with readings printed in *kana* (phonograms) and *kanji* (Chinese characters). The motivation for using a kana version instead of romanization is that if an instructed learner living in

3. The verb ending forms used in the test items are *-masu* and *-masita*, polite forms for nonpast *-ru* and past *-ta*. The learners were more familiar with this form because polite style is usually introduced earlier in classroom settings.

Japan lacks kana knowledge, it is highly likely that he or she is lacking even basic Japanese language ability and will not understand the test items on the whole. To ensure that the learners understood the meaning of the sentences, English translations were provided for some content words, which were determined to be possibly difficult based on the reactions of participants in the pilot study.[4] Both the kana version and the romanized version of a test question are illustrated in the Appendix.

3.2.2 *Procedure*

The data were individually collected in various places such as the learner's home, office, or classroom. The participants had an OPI first, and then they completed the oral picture description task[5] and the acceptability judgment test, and, finally, filled out background questionnaires.

The two tasks were not timed. The oral task took about 10 minutes to complete, and the judgment test took about 30 minutes on average. One of the researchers (the first author) was present throughout the administration. For the judgment test, the researcher told the participants that they were free to ask about unfamiliar words in the test items, in order to properly elicit their tense-aspect knowledge and avoid misunderstanding of the test sentences. When asked, the researcher explained words by paraphrasing in Japanese or by using drawings and gestures.

Shortly after the data collection, the participants received feedback on the results of the OPI and on their general Japanese skills, and were given monetary compensation.

4. Analysis and results

We collapsed two L1 groups – NSs of English, which has the obligatory progressive, and NSs of languages that have no obligatory progressive marking (German and Slavic languages) – because previous analysis (Sugaya & Shirai 2007) found no significant difference between the two groups. Then, to examine the relationship between learners' proficiency and productivity, we assigned a

4. Most participants in the L1 nonprogressive group had a good command of English and therefore had no difficulty understanding English translations.

5. The results of the oral picture description task as well as the judgment task concerning the difficulty of resultative vs. progressive meanings of *-te i-ru* and effects of learners' L1 are reported elsewhere (Sugaya & Shirai 2007).

score to each learner based on his or her appropriate judgment of simple past and nonpast for 33 items (see Table 2). Participants were divided into lower or higher levels with the median (29) as the cut-off. Table 2 shows the number of participants, the means, the standard deviations, and the range of the score for each of the groups.[6]

Table 2. Judgment test: Scores on simple nonpast and past contexts

Group	M	SD	Range
Higher (n = 27)	31.52	0.75	30–33
Lower (n = 34)	25.59	2.73	19–29

4.1 Lower proficiency learners

The first analysis compared the accuracy score for each context by the lower proficiency learners. Table 3 shows the accuracy score. The results showed huge variation in accuracy by verbs and contexts, especially in imperfective *-te i-ru* and past *-ta* contexts.

As the criterion of verb-specific preferences, we set a cut-off point of 40% difference[7] between the accuracy score of the simple past and the nonpast imperfective target. Four verbs, which are highlighted in Table 3, met this criterion. Two verbs were preferred in imperfective *-te i-ru* form. *Siru* 'come to know' and *tuku* 'be attached' showed high accuracy scores in *-te i-ru* contexts (97% for *siru* and 79% for *tuku*), but low scores in *-ta* contexts (41% for *siru* and 29% for *tuku*). Thus, learners prefer *-te i-ru* predominantly for these verbs. In contrast, *otiru* 'fall' and *iku* 'go' showed low accuracy scores in *-te i-ru* contexts (47% for *otiru* in resultative *-te i-ru*, 53% for *iku*), but high scores in *-ta* contexts (94% for *otiru*, 97% for *iku*). In other words, the learners preferred to use these verbs in past tense form. Importantly, all these verbs are in resultative contexts, while no verb in progressive contexts met this criterion. This indicates that the verb-specific preferences were more strongly observed with the verbs for the resultative state than the verbs for the progressive meaning.

6. Generally, learners classified as lower proficiency correspond to intermediate, and those classified as higher proficiency correspond to advanced, based on the OPI conducted as part of the study.

7. Even if we choose 30% as the cut-off, the trend is the same: No verb shows verb-specific preferences for the higher-proficiency group, and more verbs used for resultative *-te i-ru* met the criterion than those for progressive *-te i-ru* (5 vs. 2).

Table 3. Accuracy scores by verb forms: Lower proficiency learners

Target contexts (Correct forms)	Simple nonpast (*-ru*)	Simple past (*-ta*)	Nonpast imperfective (*-te i-ru*)	Mean
Verbs for progressive meaning with *-te i-ru*				
Activity				
asobu 'play'	85%	65%	88%	79%
huru 'rain'	94%	62%	100%	85%
nomu 'drink'	91%	88%	94%	91%
Accomplishment				
karee-o tukuru 'cook curry'	76%	100%	71%	82%
naraberu 'line (something) up'	62%	38%	76%	59%
repooto-o kaku 'write a term paper'	88%	97%	100%	95%
Semelfactive				
otiru 'fall'	94%	94%	68%	85%
tataku 'beat'	76%	91%	94%	87%
tiru '(cherry blossoms) fall'	65%	76%	68%	70%
Mean	81%	79%	84%	82%
Verbs for resultative meaning with *-te i-ru*				
Achievement				
iku 'go'	79%	97%	53%	76%
kekkon-suru 'get married'	97%	91%	76%	88%
kowareru 'break' (intr.)	76%	79%	74%	76%
oboeru 'memorize'	74%	62%	85%	74%
otiru 'fall'	94%	94%	47%	78%
siru 'come to know'	–	41%	97%	69%
todoku 'arrive'	94%	53%	18%	55%
tukareru 'get tired'	79%	68%	79%	75%
tuku 'be attached to'	91%	29%	79%	66%
Mean	86%	68%	68%	73%

4.2 Higher proficiency learners

Table 4 shows the results from the higher proficiency learners. The results indicated that the response pattern of the higher-level group was different from that of the lower group. The higher proficiency learners correctly chose tense-aspect forms, demonstrating an accuracy score of 80% or higher for almost all verbs in all three target contexts (simple nonpast, past, and nonpast imperfective), and no verb met the 40% criterion of verb-specific use discussed above.

To summarize, we found that the advanced learners correctly chose tense-aspect forms in three contexts. In contrast, the lower level learners showed verb-specific preference and it was stronger with the verbs for resultative meaning.

Table 4. Accuracy scores by verb forms: Higher proficiency learners

Target contexts (Correct forms)	**Simple nonpast** (*-ru*)	**Simple past** (*-ta*)	**Nonpast imperfective** (*-te i-ru*)	**Mean**
Verbs for progressive meaning with *-te i-ru*				
Activity				
asobu 'play'	100%	96%	100%	99%
huru 'rain'	100%	85%	96%	94%
nomu 'drink'	100%	96%	100%	99%
Accomplishment				
karee-o tukuru 'cook curry'	96%	100%	100%	99%
naraberu 'line (something) up'	85%	70%	93%	83%
repooto-o kaku 'write a term paper'	100%	100%	100%	100%
Semelfactive				
otiru 'fall'	100%	100%	85%	95%
tataku 'beat'	100%	100%	100%	100%
tiru '(cherry blossoms) fall'	100%	96%	89%	95%
Mean	98%	94%	96%	96%
Verbs for resultative meaning with *-te i-ru*				
Achievement				
iku 'go'	96%	100%	96%	97%
kekkon-suru 'get married'	100%	100%	85%	95%
kowareru 'break' (intr.)	100%	100%	96%	99%
oboeru 'memorize'	100%	96%	93%	96%
otiru 'fall'	100%	100%	89%	96%
siru 'come to know'	–	93%	100%	97%
todoku 'arrive'	96%	67%	78%	80%
tukareru 'get tired'	100%	89%	100%	96%
tuku 'be attached to'	100%	89%	96%	95%
Mean	99%	93%	93%	95%

5. Discussion

This study set out to answer the following research questions:

1. Do L2 learners use Japanese tense-aspect markers in verb-specific fashion? If so, is there an effect of verb types (i.e., verbs denoting progressive meaning with *-te i-(ru)*) vs. verbs denoting resultative state meaning with *-te i-(ru)*)?
2. Do L2 learners eventually attain productive control of Japanese tense-aspect markers?

In answer to the first research question, the results showed that the lower proficiency learners used the individual verbs in verb-specific ways and that this tendency was stronger for the verbs denoting resultative state meaning with *-te i-(ru)* (e.g., achievement verbs) than the verbs denoting progressive meaning with *-te i-(ru)* (e.g., activity, accomplishment, and semelfactive verbs).

In answer to the second research question concerning L2 development, we found that the higher-level learners showed high accuracy scores, more than 80% accuracy for almost all verbs in all three contexts (simple nonpast, past, and imperfective) with an overall accuracy score of 95.6%, which means that they eventually attained productive control of Japanese tense-aspect markers.

In what follows, we discuss theoretical implications of these findings in relation to the learning of formulaic sequences in second language acquisition. We first discuss the role of rote learning in the acquisition of tense-aspect, in particular with reference to the Aspect Hypothesis. Then, we discuss the issue concerning how rote learning and rule learning are related in the acquisition of tense-aspect marking.

5.1 Why is there more verb-specific use for resultative use of *-te i-ru*?

As discussed above, the pattern of development of tense-aspect markers has been extensively studied in relation to the Aspect Hypothesis, which predicts a strong association between past tense markers and telic verbs and of progressive marking with activity verbs. The correlation between grammatical aspect and inherent lexical aspect has been noted crosslinguistically; however, why such universal patterns are observed is still an open issue; various explanations have been proposed, such as frequency in the input (Andersen & Shirai 1996), conceptual saliency of some aspectual notions (Andersen 1991), an innate bioprogram (Bickerton 1981), and L1 transfer (Shirai & Kurono 1998; see Sugaya & Shirai 2007 for further discussion).

In the context of L2 acquisition of Japanese, the association is often translated into early acquisition of imperfective *-te i-(ru)* with progressive meaning rather than with resultative meaning, since progressive meaning is obtained when combined with activity verbs, while resultative meaning is obtained with achievement verbs, which in other languages (and also in Japanese) are strongly associated with past tense markers. Input frequency has also been invoked as an explanation; however, the difficulty here is that in Japanese, *-te i-(ru)* is more frequently used with achievement verbs in native speech than with activity verbs. Given that frequency does not work well as an explanation, the simpler form-meaning relationship has been proposed as an explanation for the early acquisition of progressive meaning of *-te i-(ru)*; that is, progressive meaning has no competitor in Japanese, while resultant state can also be expressed using past tense form *-ta* (Shirai 1993; Sugaya & Shirai 2007).

This complex form-meaning mapping for resultative *-te i-(ru)* may be an important reason why learners showed more verb-specific patterns for achievement verbs. In fact, nine verbs denoting progressive with *-te i-ru* (three activities, 85%; three accomplishments, 79%; and three semelfatives, 81%) have an overall higher mean accuracy score (82%) than those verbs denoting resultative *-te i-ru* (73%) for lower proficiency learners (see Table 3). This difference, importantly, disappears for higher proficiency learners, who can handle both types of meanings effectively. Following the argument advocated earlier, we speculate that since it is more difficult to map form to meaning in the case of achievement verbs with *-te i-(ru)*, learners have to rely more on frequency-based rote learning. That is, since there is no competition in referring to progressive meaning combined with activity verbs, learners can rely on rule-based learning for progressive meaning, while it is more difficult to make form-meaning mappings for resultative *-te i-(ru)* due to the major competing form (past *-ta*), so that learners have to rely on item-based learning. DeKeyser (1995) showed that rule learning is easier for categorical rules, whereas exemplar-based learning is more effective for non-categorical, prototype-based rules. Although there is not a complete parallel here, it is certainly possible to extend DeKeyser's idea to Japanese *-te i-(ru)*, since progressive meaning has simpler (i.e., more categorical) form-meaning mapping than resultative *-te i-(ru)*.

5.2 Distributional bias: What kind?

Another issue is what kind of frequency information learners attend to in making form-meaning connections. Recall that some achievement verbs (*iku, otiru*) in this study are associated with past tense *-ta*, congruent with the Aspect Hypothesis, while others (*siru*, *tuku*) are associated with imperfective *-te i-ru*, going against the Aspect Hypothesis. It is noteworthy that Alla (Sugaya 2003, 2005), the learner discussed earlier, showed the similar verb-specific preference for *siru*, *tuku* (associated with *-te i-ru*), and *iku* (associated with *-ta*). (She did not use the verb *otiru*.)[8]

Why did some achievement verbs, contra the prediction of the Aspect Hypothesis, not show the preference for past tense form? A natural explanation is to attribute this to a frequency effect. Take, for example, the verb *siru* 'come to know,' which intermediate learners preferred to use in the *-te i-ru* form. Even

8. Analysis of individual response shows that this verb-specific preference is quite robust. For the four verbs, the pattern of errors follows the verb-specific preferences observed in the group data. For example, although there are 19 learners who got either *-ta* or *-te i-ru* correct with *siru* 'know' (but not both), none of them got *-ta* correct (0 vs. 19). Similarly, for *tuku* it was 3 vs. 20 in favor of *-te i-ru*. In contrast, *-ta* preferred verbs showed the opposite pattern in favor of *-ta*: *otiru* (17 vs. 1) and *iku* (16 vs. 1).

though we can use the verb *siru* in any of the four forms in real-life language use, the *-te i-ru* form appears to be the most frequent because we use the *-te i-ru* form (*sit-te i-ru*) to refer to the present state of knowing something/someone (e.g., *Ken-wa Naomi-o sit-te i-ru* 'Ken knows Naomi').

To examine the frequency of use by native Japanese speakers, we used the demo version of the KOTONOHA corpus, which has been recently released by the National Institute for Japanese Language (http://www.kotonoha.gr.jp/demo/). The corpus was randomly collected from two sources: white papers from the Japanese government (1500 texts, five million words) and *Yahoo Chiebukuro* 'Yahoo! Answers in Japanese,' which provides online knowledge community service in question-and-answer format (45725 texts, five million words). The demo version comes with a simple concordance program, but it only shows up to 500 examples of the target item even when there are more target items. Nonetheless, the corpus provides us with informative data relevant to our study, and we bypassed this problem by choosing a smaller sample (*Yahoo Chiebukuro* 2004, instead of 2005) so that there would not be more than 500 hits in one search. We chose *Yahoo Chiebukuro* rather than white papers for obvious reasons: it deals with everyday topics such as shopping, housing, cooking, computers, travel, weather, job-related matters, and so on, and we assumed it more closely matches everyday language use, to which L2 learners residing in Japan are exposed.

Table 5 illustrates the frequency of each tense-aspect form for all achievement verbs used in the judgment test, for which the lower level learners showed verb-specific preferences. First, let us look at the four verbs that showed verb-specific tendency (*siru* 'come to know' and *tuku* 'be attached to,' for which learners preferred imperfective *-te i-ru* form, and *otiru* 'fall' and *iku* 'go,' for which learners preferred past - *ta* form). *Siru* 'come to know' was used predominantly with *-te i-ru* (73%), which explains why even the lower proficiency learners showed a high accuracy score (97%) for *siru* in the resultative *-te i-ru* target, even though previous studies found that resultative state meaning is generally more difficult than progressive meaning. The case of *tuku* 'be attached' is not so straightforward since frequency of nonpast form (41%) is higher than *-te i-ru* form (32%), but considering that nonpast *-ru* form is the unmarked form of the verbs, the frequency used with *-te i-ru* is noteworthy.

When we look at the verb that had high accuracy with past tense *-ta* over imperfective *-te i-ru* (*-ta* preferred verbs in Table 5), for both *otiru* 'fall' and *iku* 'go,' the simple nonpast form was the highest in frequency (about 60%) in the same corpus, followed by past, nonpast imperfective, and past imperfective. Therefore, if we look only at the four verbs that showed verb-specific preferences, we can come up with a tentative conclusion that the preference follows the relative frequency of past *-ta* vs. nonpast imperfective *-te i-ru*; whichever is more frequent

Table 5. Distribution of the four tense-aspect forms from *Yahoo! Chiebukuro* 2004

	Simple nonpast	Simple past	Nonpast imperfective	Past imperfective
	(*-ru*)	(*-ta*)	(*-te i-ru*)	(*-te i-ta*)
-te i-ru preferred verbs				
siru 'to come to know'	61 (11%)	67 (12%)	394 (73%)	18 (3%)
tuku 'to be attached to'	262 (41%)	140 (22%)	208 (32%)	31 (5%)
-ta preferred verbs				
otiru 'to fall'	81 (59%)	36 (26%)	16 (12%)	4 (3%)
iku 'to go'	788 (64%)	354 (29%)	75 (6%)	22(2%)
No preference verbs				
kekkon-suru 'get married'	79 (44%)	75 (42%)	22 (12%)	4 (2%)
kowareru 'break' (intr.)	26 (39%)	21 (31%)	19 (28%)	1 (1%)
oboeru 'memorize'	29 (31%)	10 (11%)	52 (56%)	2 (2%)
todoku 'arrive'	88 (43%)	83 (41%)	28 (14%)	4 (2%)
tukareru 'get tired'	42 (49%)	21 (24%)	19 (22%)	4 (5%)

Note. All finite verb tokens in both the matrix clause and the subordinate clause were included except for the negative forms.

determines the verb-specific preferences. Considering the fact that nonpast *-ru* is an unmarked form, this may be a reasonable hypothesis.[9]

The picture may not look that simple once we look at the verbs that did not show any verb-specific preferences. Specifically, there is a strong frequency bias for *kekkon-suru* 'get married,' *todoku* 'arrive' (for past *-ta*), and *oboeru* 'memorize' (for *-te i-ru*). However, the results from the judgment test actually follow the frequency bias. Though they did not pass the 40% criteria we set, all these verbs have a higher accuracy score for the frequent form: Past is more accurately judged for *kekkon-suru* (by 15%) and *todoku* (by 35%), while *-te i-ru* is more accurately

9. This hypothesis is consistent with the frequency data in native Japanese speech reported in Andersen (1993), which reports two studies. In both studies, past *-ta* is strongly associated with achievement verbs (about 50%) while nonpast *-ru* is associated with activity in one study (65%), but with state (38%) and achievement (33%), suggesting that *-ru* does not have any strong association with a particular lexical aspect class. (The numbers are based on token count).

judged for *oboeru* (by 23%) (see Table 3). For the two remaining verbs for which frequency differences were minimal, the differences in accuracy were also small (5% for *kowareru* 'break' and 11% for *tukareru* 'get tired'). Thus, it is clear that there is a strong effect of frequency in terms of verb-specific preferences exhibited by the learners for these achievement verbs.

In order to further investigate the effect of frequency, we calculated Spearman's Rho between the accuracy score in the judgment test and the frequency for these verbs (both percentages). Interestingly, correlation for nonpast imperfective *-te i-ru* was highest and statistically significant ($rs = .803$, $p < .05$), while others were not: past *-ta* ($rs = .483$, *ns*), nonpast *–ru* ($rs = .446$, *ns*). These correlational coefficients make sense. As noted above, *-ru* being an unmarked form, its relation with particular forms should not be strong. Regarding the past tense *-ta*, it may not be as sensitive to verb-specific frequency since learners can make simpler form-meaning association at the level of meaning (i.e., past form is most frequently used with achievement verbs). Since *-te i-(ru)* is frequently used both with achievements (to denote resultative state) and activity (to denote progressive) learners must be more sensitive to which verbs are used in what form. This results in stronger reliance on rote learning and hence higher correlation with frequency information and accuracy scores in the judgment test.

The results of the present study also have an important implication for the Aspect Hypothesis. Although effect of input frequency has been discussed in relation to the Distributional Bias Hypothesis (Andersen 1993; Andersen & Shirai 1994), that research only considered the level of correlation of verb morphology and verb classes. That is, it did not consider the effect of frequency at the level of each individual verb. In fact, as the present study shows (see also Shiokawa 2006), some achievement verbs are more associated with the imperfective *-te i-(ru)*, contra the Aspect Hypothesis, which predicts their association with past tense *-ta*. Further research is needed to investigate the relative contribution of form-meaning association based on semantic generalization and form-form association, which is purely syntagmatic.

5.3 Verb-specific pattern vs. rule-based learning in L2 acquisition of tense-aspect

The present study shows that early acquisition of Japanese tense-aspect morphology shows verb-specific patterns, and that learners then attain productive control of tense-aspect forms. This is consistent with the idea of developmental progression proposed by N. Ellis (2002) and Tomasello (2003). Here we hypothesize the following process of development of tense-aspect markers in Japanese as a second language: from formulaic, rote-learned forms at the elementary level, as in the

case of Alla, whose OPI level was mostly at elementary level,[10] and then to low-scope patterns at the intermediate level and finally to productive constructions at the advanced level.

Formula	>	low-scope pattern	>	constructions
Elementary		Intermediate		Advanced
(Alla)		(present study)		(present study)

What is the implication of the results for the issue of the dual-process model we touched upon at the beginning of the paper? The results are perfectly consistent with the position that item-based formulaic learning serves as the basis for creative language learning. The present study is also important in that it used a judgment test, which was less likely to burden learners with communicative pressure, hence facilitating the use of rote-learned formulaic expressions (Krashen & Scarcella 1978). Even with such data, learners showed verb-specific preferences, which indicate that item-based learning may contribute to learners' knowledge of semantics.

Having said that, we must admit that we cannot rule out the possibility that rote-learned process and rule-learning are separate processes, as proposed by Krashen and Scarcella (1978), because it is almost impossible to test this by looking at behavioral data only, given that there is no telling which process is at work. Further complicating the issue, it has been proposed that memory-based and rule-based processes co-exist for particular linguistic items (Langacker 1987; Bybee 1985), and that linguistic knowledge should be considered a "formulaic-creative continuum" (e.g., Bolinger 1976; Goldberg 2003). Nevertheless, the issue of rule vs. memory is an important issue in cognitive science (e.g., McClelland & Patterson 2002; Pinker & Ullman 2002). We probably need more data from neuroscience to address the issue of how learners represent these dual processes in their learning and use of second languages.

6. Conclusion

The results from the judgment test in this study suggest that the early acquisition of Japanese tense-aspect morphology shows verb-specific patterns and that learners gradually attain productive control of tense-aspect forms, which is consistent with the proposed developmental sequence of formula > low-scope pattern > construction (N. Ellis 2002; Tomasello 2003). The asymmetry between progressive and

10. Of course, this does not mean that Alla used only formulas. For some verbs, she showed productive control of tense-aspect morphology as well. It is just that her use of tense-aspect markers was mostly formulaic.

resultative meaning will be accounted for by multiple factors, such as frequencies of imperfective forms in native Japanese speech and complexity of form-meaning mapping (Sugaya & Shirai 2007).

The present study also has implications for studies of the acquisition of aspect in other languages. As noted above, most studies on aspect focus on association between tense-aspect marking and verb semantics. However, as the present study shows, there may be verb-specific association within particular lexical aspect class, on which frequency-based rote learning appears to have a strong effect. Investigation into verb-specific preference within lexical aspect class would be an important area in our full understanding of tense-aspect acquisition, in particular regarding the role of rote vs. rule learning and how they interact in second language acquisition.

Appendix

Sample test item

(1) Original *Kana* version
高橋：あれ、シャツに口紅 (lipstick) が＿＿＿＿＿＿ね。
山本：え、ほんとうですか！？
A. つきます　B. つきました　C. ついています　D. ついていました

(2) Romanized version and gloss
Takahasi: Are, syatu-ni kutibeni (lipstick)-ga ＿＿＿＿＿＿ne
Oh, shirt-LOC lipstick-NOM ＿＿＿＿＿＿–FP.
'Oh, there's lipstick on your shirt.'
Tanaka: E, hontoo desu-ka!?
Oh, true-COP-Q!?
'Oh, really!?'

A. tukimasu	B. tukimasita	C. tuiteimasu	D. tuiteimasita
attach: NONPAST	attach: PAST	attach: IPFV-NONPAST	attach: IPFV-PAST
'is attached'	'was attached'	'has been attached'	'had been attached'

Note. The correct answer is C for this test item. The verb ending forms used in the test items are *-masu* and *-masita*, polite forms for nonpast *-ru* and past *-ta*.

References

Andersen, Roger W. 1991. Developmental sequences: The emergence of aspect marking in second language acquisition. In *Crosscurrents in second language acquisition and linguistic theories,* T. Huebner & C.A. Ferguson (Eds), 305–324. Amsterdam: John Benjamins.

Andersen, Roger W. 1993. Four operating principles and input distribution as explanations for underdeveloped and mature morphological systems. In *Progression and regression in language,* K. Hyltenstam & Å. Viborg (Eds), 309–339. Cambridge: CUP.

Andersen, Roger W. & Yasuhiro Shirai. 1994. Discourse motivations for some cognitive acquisition principles. *Studies in Second Language Acquisition* 16(2): 133–156.

Andersen, Roger W. & Yasuhiro Shirai. 1996. The primacy of aspect in first and second language acquisition: The pidgin-creole connection. In *Handbook of second language acquisition*, W.C. Ritchie & T.K. Bhatia (Eds), 527–570. San Diego CA: Academic Press.

Bardovi-Harlig, Kathleen. 1999. From morpheme studies to temporal semantics: Tense-aspect research in SLA. *Studies in Second Language Acquisition* 21(3): 341–382.

Bardovi-Harlig, Kathleen. 2002. A new starting point? Investigating formulaic use and input in future expression. *Studies in Second Language Acquisition* 24(2): 189–198.

Bickerton, Derek. 1981. *Roots of language.* Ann Arbor MI: Karoma.

Bohn, Ocke-Schwen. 1986. Formulas, frame structures, and stereotypes early syntactic development: Some new evidence from L2 acquisition. *Linguistics* 24: 185–202.

Bolinger, Dwight. 1976. Memory and meaning. *Forum Linguisticum* 41(1): 1–14.

Brown, Roger. 1973. *A first language*. Cambridge MA: Harvard University Press.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form.* Amsterdam: John Benjamins.

Clark, Ruth. 1974. Performing without competence. *Journal of Child Language* 1(1): 1–10.

DeKeyser, Robert. M. 1995. Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition* 17(3): 379–410.

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

Ellis, Rod. 1984. Formulaic speech in early classroom second language development. In *On TESOL '83: The question of control,* J. Handscombe, R.A. Orem & B.P. Taylor (Eds), 53–65. Washington DC: TESOL.

Ellis, Rod. 1999. Item versus system learning: Explaining free variation. *Applied Linguistics* 20(4): 460–480.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5): 219–224.

Hakuta, Kenji. 1974. Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning* 24(2): 287–297.

Koyama, Satoru. 2003. Nihongo no tensu-asupekuto no syuutoku ni okeru huhensei to kobetusei: Bogo no yakuwari to eikyoo o tyuusin ni (Universality and variability in the acquisition of Japanese tense-aspect: Focusing on the role and effect of L1). In *Gengo to kyooiku: Nihongo o taisyoo ni* (Language and education: The case of Japanese), S. Koyama, K. Otomo & M. Nohara (Eds), 415–436. Tokyo: Kurosio.

Krashen, Stephen D. & Robin Scarcella. 1978. On routines and patterns in language acquisition. *Language Learning* 28(2): 283–300.

Kurono, Atsuko. 1998. Ryuugakusei no hatuwa ni mirareru tensu-asupekuto no goyou ni tuite (Inappropriate use of tense-aspect markers by L2 learners of Japanese). *Final report for the grant-in-aid for scientific research* (Basic research on the development of speaking ability and environment of use in Japanese by researchers from overseas), 112–124. PI: Akito Ozaki, Nagoya University.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar: Theoretical prerequisites.* Stanford CA: Stanford University Press.

Li, Ping & Yasuhiro Shirai. 2000. *The acquisition of lexical and grammatical aspect*. Berlin: Mouton de Gruyter.

McClelland, James L. & Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6(1): 465–472.

Myles, Florence, Rosamond Mitchell & Janet Hooper. 1999. Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition* 21(1): 49–80.

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

Pinker, Steven & Michael T.Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences* 6(11): 456–463.

Robison, Richard. E. 1995. The aspect hypothesis revisited: A cross-sectional study of tense and aspect marking in interlanguage. *Applied Linguistics* 16(3): 344–370.

Sheu, Shiapei. 2005. *Nihongo gakusyuusya ni yoru asupekuto no syuutoku* (A study of the acquisition of aspect marker by learners of Japanese). Tokyo: Kurosio.

Shibata, Miki. 1999. The use of Japanese tense-aspect morphology in L2 discourse narratives. *Acquisition of Japanese as a Second Language* 2: 68–102.

Shibata, Miki. 2000. Function of tense-aspect morphemes in second language discourse. Ph.D. dissertation, University of Arizona, Tucson.

Shiokawa, Eriko. 2006. Nihongo gakusyuusya no asupekuto keisiki no syuutoku (A study on the acquisition of aspect form by L2 learners of Japanese). Paper presented at the International Symposium on Japanese Language Education, Chinese University of Hong Kong, October, 29–30.

Shirai, Yasuhiro. 1991. Primacy of aspect in language acquisition: Simplified input and prototype. Ph.D. dissertation, University of California, Los Angeles.

Shirai, Yasuhiro. 1993. Inherent aspect and acquisition of tense/aspect morphology in Japanese. In *Argument structure: Its syntax and acquisition*, H. Nakajima & Y. Otsu (Eds), 67–82. Tokyo: Kaitakusya.

Shirai, Yasuhiro. 1995. Tense-aspect marking by L2 learners of Japanese. In *Proceedings of the 19th Annual Boston University Conference on Language Development,* Vol. 2, D. MacLaughlin & S. McEwen (Eds), 575–586. Somerville MA: Cascadilla.

Shirai, Yasuhiro. 2000. The semantics of the Japanese imperfective *-te i-ru*: An integrative approach. *Journal of Pragmatics* 32(3): 327–361.

Shirai, Yasuhiro. 2004. A multiple-factor account for the form-meaning connections in the acquisition of tense-aspect morphology. In *Form-meaning connections in second language acquisition*, B. VanPatten, J. Williams, S. Rott & M. Overstreet (Eds), 91–112. Mahwah NJ: Lawrence Erlbaum Associates.

Shirai, Yasuhiro & Roger W. Andersen 1995. The acquisition of tense-aspect morphology: A prototype account. *Language* 71(4): 743–762.

Shirai, Yashiro & Atsuko Kurono. 1998. The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning* 48(2): 245–279.

Shirai, Yasuhiro & Yumiko Nishi. 2005. How what we mean impacts how we talk: The Japanese imperfective aspect marker *-te i-ru* in conversation. In *The power of context in language learning and teaching*, J. Frodesen & C. Holten (Eds), 39–48. Boston MA: Thomson Heinle.

Smith, Carlota S. 1991. *The parameter of aspect*. Dordrecht: Kluwer.

Sugaya, Natsue. 2003. Nihongo gakusyuusya no asupekuto syuutoku ni kansuru zyuudan kenkyuu: Doosa no zizoku to kekka no zyootai no *-te i-ru* o tyuusin ni (A longitudinal study on the acquisition of imperfective aspect morphology by L2 learners of Japanese: Focusing on the progressive and the resultative state use of *-te i-ru*). *Nihongo Kyooiku* (*Journal of Japanese Language Teaching*) 119: 65–74.

Sugaya, Natsue. 2005. Dainigengo tositeno nihongo no asupekuto syuutoku kenkyuu: Naizai asupekuto to bogo no yakuwari (A study on the acquisition of the imperfective aspect marker in Japanese as a second language: The role of inherent aspect and L1). Ph.D. dissertation, Ochanomizu University.

Sugaya, Natsue & Yashiro Shirai. 2007. The acquisition of progressive and resultative meanings of the imperfective aspect marker by L2 learners of Japanese: Universals, transfer, or multiple factors? *Studies in Second Language Acquisition* 29(1): 1–38.

Tomasello, Michael. 1992. *First verbs: A case study of early grammatical development*. Cambridge: CUP.

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

Uozumi, Tomoko. 1998. Tuiseki tyoosa ni mirareru *-te i-ru* no syuutoku zyookyoo (The acquisition of *-te i-ru* in the follow-up study). *Final report for the grant-in-aid for scientific research* (Basic research on the development of speaking ability and environment of use in Japanese by researchers from overseas), 100–111. PI: Akito Ozaki, Nagoya University.

Vendler, Zeno. 1967. *Linguistics in philosophy*. Ithaca NY: Cornell University Press.

Vihman, Marilyn M. 1982. Formulas in first and second language acquisition. In *Exceptional language and linguistics*, L.K. Obler & L. Menn (Eds), 261–284. New York NY: Academic Press.

Weinert, Regina. 1995. The role of formulaic language in second language acquisition: A review. *Applied Linguistics* 16(2): 180–205.

Weist, Richard M. 2002. The first language acquisition of tense and aspect: A review. In *The L2 acquisition of tense-aspect morphology*, R.M. Salaberry & Yasuhiro Shirai (Eds), 21–78. Amsterdam: John Benjamins.

Wode, Henning. 1981. *Learning a second language*. Tübingen: Narr.

Wong-Fillmore, Lily. 1976. The second time around: Cognitive and social strategies in second language acquisition. Ph.D. dissertation, Stanford University.

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

Formulaic and novel language in a 'dual process' model of language competence

Evidence from surveys, speech samples, and schemata

Diana Van Lancker Sidtis
New York University
Nathan Kline Institute

Abstract

The fact that formulaic expressions are consistently preserved in left hemisphere damage has had little influence on models of language. Evidence from disordered speech, linguistic analyses, and first and second language learning reveals that formulaic and novel expressions pattern differently. The "formuleme" (canonical form) is recognizable by native speakers as having stereotyped form and conventional meaning. Studies suggest that one quarter of discourse is made up of formulaic expressions, and that right hemisphere and subcortical damage interfere with their comprehension and production. The dual process model features a holistic mode for processing of formulaic language and an analytic mode for generation of new utterances. Schemata (formulemes with open slots) exemplify normal cooperation between generation of fixed and newly created language.

1.1 Background

My attraction to formulaic language arose not out of any neat linguistic insight, but rather from exposure to aphasic speech. When first observing speech therapy sessions in rehabilitation centers around Los Angeles, it became apparent that persons with language difficulties, even very severe ones, while struggling and failing to talk in standard ways, fluently produced certain kinds of speech with normal articulation and prosody. A literature review revealed that knowledge of preserved speech in aphasia appeared in virtually every clinical description since the mid nineteenth century, usually indexing similar phenomena with overlapping categories (Espir & Rose 1970; Goodglass & Kaplan 1972; Van Lancker 1975, 1988, 1993). While the terminology was inconsistent, the ubiquity of "automatic speech" commentary in the earlier clinical literature can hardly be exaggerated (Alajouanine 1956; Bay 1964; Benson 1979; Critchley 1970; Gloning, Gloning & Hoff 1963; Goldstein 1948; Goodglass & Mayer 1958; Head 1926; Luria 1964, 1966; Pick 1973). The categories include serial speech (such as counting), memorized expressions, sayings, nursery rhymes, familiar lyrics, prayers, clichés, yes, no, greetings and salutations, onsets of sentences ("I want, I can") as well as idiosyncratic recurrent utterances in individual patients' repertories.

The most well-known and influential of the early writers on aphasia, the neurologist John Hughlings Jackson (1874), provided vivid examples of preserved aphasic speech, and elaborated a brain model that differentiated what he termed "propositional" and "automatic" (or "nonpropositional") speech. In Jackson's formulation, these are natural human abilities associated with left and right hemisphere processing respectively, and are differentially affected by brain damage (Van Lancker 1975). The celebrated example from Baudelaire, the great French poet, who suffered a left hemisphere stroke at age 45, was well known: his only remaining utterance was "Cré nom," part of a French curse (Dieguez & Bogousslavsky 2007). Although definitions and details have evolved and changed somewhat, these ideas remain pertinent and modern today. Yet despite the resilience and accuracy of the notion that some types of speech are dramatically unaffected by brain damage causing language disturbance, none of these ideas had found their way into linguistic models of language competence (Van Lancker 1973).

2.1 Definitions and description

A considerable range of expressions can be categorized as nonpropositional, using the criterion that they are not novel – that is, they are not newly created

from the operation of grammatical rules on lexical items[1] (Figure 1). These include idioms[2] proverbs, speech formulas, conventional expressions, expletives, and so on. Besides being not newly created from units (lexical and morphological elements) and rules, they have other characteristics in common: stereotyped form, conventionalized meaning, and familiarity. Stereotyped form means that formulaic expressions contain precisely specified words in a certain word order spoken on a set intonation contour. Secondly, the meanings are conventionalized, which means they are idiosyncratic in various ways, either by being nonliteral, or serving mainly as social signals, or merely by, as Wray (2002) has emphasized, communicating a meaning that is greater than the sum of their parts – the special innuendos. Take the expression, spoken by a co-ed to her friend, "I met someone." On the face of it, this utterance can be declarative, literal, and informational. But as a formula it has stereotyped form, including prosodic contour (accent on met, overall declination, and distinctive light, low voice quality), and in its meaning, it has innuendos of excitement and romance, which extend over and above the words themselves. (Try this example out on a college class. Students smile on hearing the utterance. This is the "smile test" for identifying formulaic expressions.)[3]

Alongside the stereotyped form and conventionalized meaning of formulas, there is also considerable flexibility, which means that many variants can and do appear. Linguists and psycholinguists have spent much energy in trying to find generalizations underlying these variations, with many conflicting claims (Van Lancker Sidtis 2006a). One approach is to consider the formula as having a canonical form (the "formuleme"), and that any alteration conforming to grammatical possibilities in the language is possible, as long as the canonical form remains recognizable. Finally, as alluded to above and often revealed by the smile test, a key feature of formulaic expressions is their familiarity: people know them. Their status as common knowledge in a linguistic community forms part of their raison d'être.

1. The reader is referred to novel sentences as described by S. Pinker: '… virtually every sentence that a person utters or understands is a brand-new combination of words, appearing for the first time in the history of the universe' (1995: 22).

2. Corresponding terms in German in preparation for examining German aphasic speech are listed in Appendix I.

3. That formulaic and novel meanings on ambiguous utterances are differently articulated and intoned can be detected from the acoustic information alone without other contextual cues by native listeners (Van Lancker, Canter, and Terbeek 1981; Van Lancker Sidtis 2003).

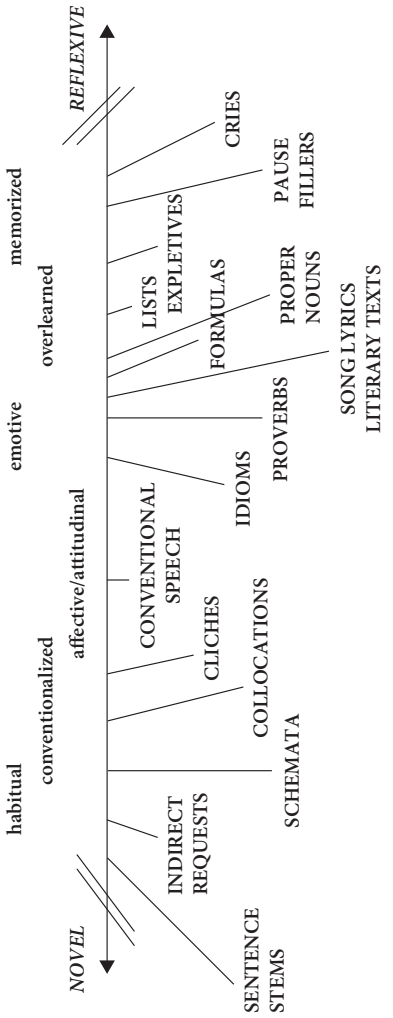


Figure 1. An heuristic continuum of nonnovel expressions.

Questions that have arisen in the course of studying formulaic language are the following:

1. How many are there? (That is, how can we figure out a way to count them?)
2. Do people know them? (That is, can we show that people know them?)
3. Are different types of formulaic expressions mentally acquired, stored, and processed differently from novel expressions, and from each other?

2.2 How many are there?

The question of how many formulaic utterances are normally used in communicative behavior engaged the interest of students at Carleton College[4] in 1998; and later at New York University. Students brought to class examples from conversations with their peers. We collated lists. As had been maintained when Chuck Fillmore engaged in a similar activity decades before at Berkeley, CA, no upper limit in numbers of formulaic expressions was seen (Fillmore 1979). Questions to ask, for example, are "How many formulas are uttered in a standard conversational interaction?" or "How many are used in the course of one day?"

At Carleton College, we chose to investigate speech behavior in a movie. Given a chance to nominate films, some students wanted to see the most recent Mike Myers production, but a more classic film made the final cut. As a classroom activity, we purchased snacks and rented a videotape to spend an evening watching "Some like it hot" (Wilder 1959) with the charge that everyone write down any and all formulaic utterances, which were collated for a total listing and count. For a rough estimate, we divided the total number into the length of the film, and were surprised that the dialogue contained a rather high rate: four formulaic expressions per minute.[5] Later, at New York University, the published screenplay (Wilder & Dimond 1959) was discovered, allowing a more leisurely examination of the dialogue. As part of this project, methods for identifying and classifying formulaic utterances in actual usage were developed in our student research group at New York University (Van Lancker Sidtis & Rallon 2004). Again, we were surprised by the large proportion: a full quarter of the utterances fell into our formulaic categories of speech formula, idiom, proverb, and conventional expression (Figure 2a).

4. Linguistics Program, Mike Flynn, Director, Northfield, MN.
5. More thorough analysis from the screenplay yielded 4.3 formulaic utterances per minute.

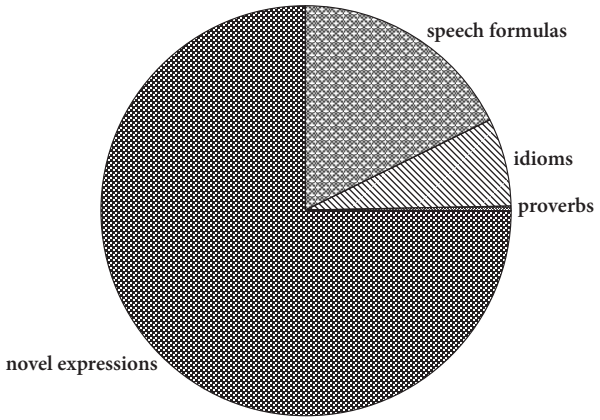


Figure 2a. Incidence of formulaic expressions in "Some like it hot."

2.3 How can we show that people know formulaic expressions?

To probe the familiarity parameter, a survey was designed, using formulaic and novel utterances randomly selected from the screenplay, to ask whether people endorse knowledge of the utterances identified by us as formulaic. In a cloze procedure, in which formulaic and novel sentences were randomized in a list, subjects performed a recall task, (entering a missing word), and a recognition task (circling "F" for familiar and "N" for novel). Subjects significantly more often provided predicted words for formulaic than novel expressions, and they also recognized both formulaic and novel expressions at a high rate (Figure 2b). This indicated that most subjects knew verbatim the majority of formulaic utterances, and that they could successfully distinguish formulaic from novel utterances.

2.4 Are they processed differently? Neurological localization of automatic speech

The question of whether formulaic expressions are processed differently from novel language can be addressed by examining neurolinguistic studies. For aphasic speech, the first steps beyond anecdotal clinical descriptions of preserved utterances, so prevalent in the aphasiological literature, were taken in England by Chris Code (1982), and in Germany by Gerhard Blanken and colleagues (1991; Blanken & Marini 1997). Speech pathologists and logopedists completed surveys

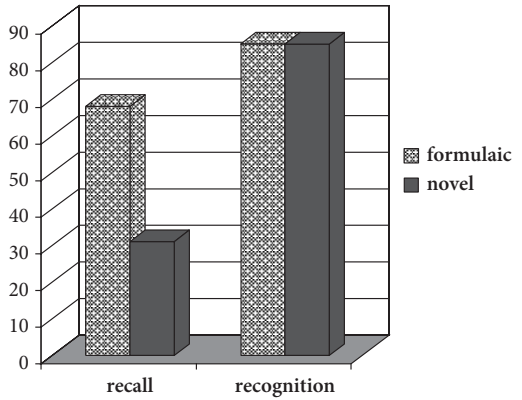


Figure 2b. Results from survey study.

providing detailed information about residual speech in severely aphasic patients. The utterances were gathered and arranged into categories, which, in both English and German, included expletives, sentence stems (e.g., I want to; *ich bin*), speech formulas (all right; *natürlich*) proper nouns, and numbers. A later analysis of residual speech in Chinese aphasic persons yielded some of these same similar utterance types (Chung, Code, and Ball 2004). This study provided documentation, classification, and theoretical consideration to preserved utterances in aphasia, and highlighted the similarity of utterance-types across individual patients and languages.

Where are these utterances represented in the brain? Aphasia is associated almost exclusively[6] with left hemisphere damage in the distribution of the middle cerebral artery, which extends over most of each hemisphere, excluding only a narrow strip on the anterior frontal lobe and another narrow area on the posterior parietal lobe. With consistent reports of preserved "subsets" of speech performance across a vast range of left hemisphere lesion sites, it seemed likely, as Hughlings Jackson (1874) had maintained, that the right hemisphere was accountable. This was not a palatable notion to many people, because current opinion held the right hemisphere to be incapable of any linguistic production. If a right hemisphere substrate were accountable, then one might expect right hemisphere damage to interfere with production of formulaic expressions.

6. Aphasia following right hemisphere damage occurs but it is extremely rare and not well understood (Basso 2003).

To address this question, Whitney Postman[7] and I examined written transcripts provided by Guila Glosser[8] of the spontaneous speech of patients who had suffered left or right hemisphere damage, as well as demographically matched normal -control subjects, speaking in comparable communicative settings (describing family and work). The method developed in the analysis of "Some like it hot" (SLIH) was expanded to cover nine categories: (1) idioms (e.g., "lost my train of thought"); (2) conventional expressions (e.g., "as a matter of fact"); (3) conversational formulaic expressions (e.g., "first of all," "right"); (4) expletives (e.g., "damn"); (5) sentence stems (e.g., "I guess"); (6) discourse particles (e.g., "well"), and (7) pause fillers (e.g., "uh"); (8) numerals; and (9) personally familiar proper nouns. While in SLIH we utilized a measure "proportion of total utterances," in the patient data the measure was changed to "proportion of words in formulaic expressions" compared to the total word count.[9] The results indicated that persons with left hemisphere damage use significantly more formulaic utterances, while persons with right hemisphere damage use significantly less, than normal subjects, rather compellingly implicating a role of the right hemisphere in production of formulaic expressions (Figure3a) (Van Lancker Sidtis & Postman 2006).

This conclusion was supported by the finding by Graves & Landis (1985) on mouth asymmetries in aphasic speakers, in which greater right sided openings (controlled by the left hemisphere) were measured for propositional tasks, while larger right sided mouth openings were observed for "automatic" tasks (e.g., counting).

To address a question posed earlier–whether subtypes of formulaic expressions, as distributed along the continuum in Figure 1, differ among themselves in neurological representation–counts for separate categories were examined. Unfortunately, in this setting, subject numbers and incidence counts were too low to draw firm conclusions. A suggestive finding was that the speech samples of the right hemisphere-damaged group contained fewer speech formulas than the other groups, and contained almost no pause fillers (Figure 3b). We are cautious because this work was based on transcripts, for which the audiotaped material was no longer available, and while we believed them to be accurate, it was possible that not every "um" and "uh" had been faithfully transcribed.

7. Studies performed at the National Institutes on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, MD.

8. Guila Glosser, Ph.D. (1951–2003), formerly at the University of Pennsylvania School of Medicine, Philadelphia, PA, tragically passed away early in the course of this project. We are grateful for her contribution.

9. Proportions of words in formulaic expressions out of the total corpus word count was utilized because a count of the total number of expressions (clauses, propositions, or sentences) is more difficult to establish in normal speech (than it was in the screenplay).

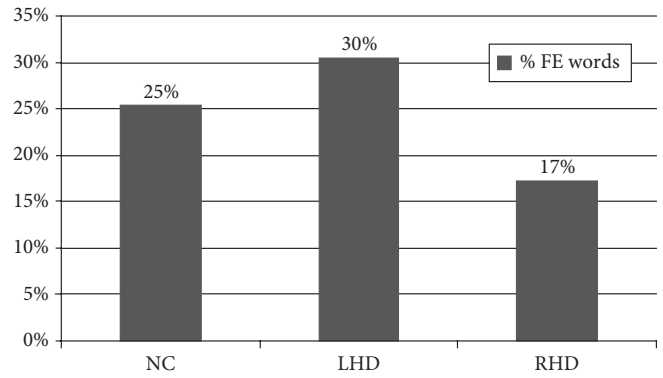


Figure 3a. Proportion of formulaic expressions in normal-control subjects, left- and right-hemisphere damaged patients.

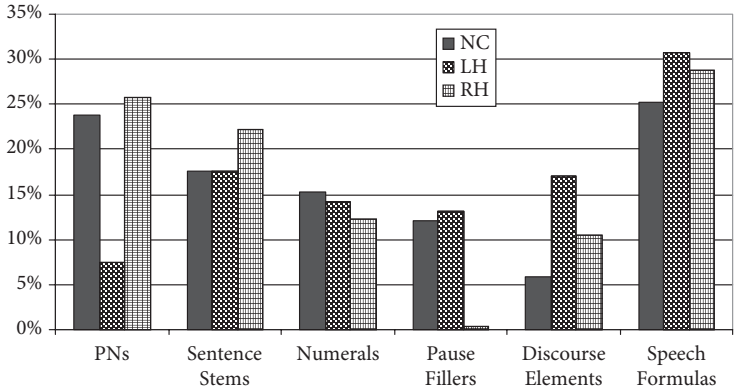


Figure 3b. Subtypes of formulaic expressions in three groups.

The next logical step was to transcribe speech from audio and videotaped material, which could be verified whenever necessary, and to perform similar analyses of formulaic expressions. Three patients, for whom extensive radiographic materials as well as language and cognitive testing were available, were studied. Case one sustained a large right hemisphere lesion, and although language abilities were intact, his conversational speech was often pragmatically inappropriate. Case two suffered right-sided subcortical damage, and like Case one, her language abilities were normal, but pragmatic elements of conversation were abnormal.

This individual complained that she no longer produced the "little words" in conversational interaction, having difficulties with greeting and leave taking. Interest in this patient was sparked by a case study by Speedie, Wertman, T'air, and Heilman (1993), describing a loss of formulaic speech production abilities following a right caudate stroke, and by previous experience with a speech disorder involving an intrusive syllable following a probable subcortical stroke. The intrusive syllable (*sis*) occurred with greater frequency during recitation, counting, and other formulaic expressions than in novel speech (Van Lancker, Bogen & Canter 1983).

The third case was a left hemisphere-damaged patient with the diagnosis of transcortical sensory aphasia (Berthier 1999; Van Lancker Sidtis 2001), who spoke fluently but with copious use of formulaic expressions. For example, when asked about his line of work, he answered "I came, I saw, I conquered." His naming response when presented with a pencil was "Reading, writing, and arithmetic." These formulaic expressions were produced with normal articulation and intonation, and considerable social confidence, such that recognition of his severe language disorder was delayed by clinical caregivers. With extensive speech samples and full background information on these three subjects, we proceeded to test hypotheses about right hemisphere and subcortical roles in production of formulaic language.

Our first concern was to develop an appropriate normal-control speech sample. Previous experience with various kinds of speech samples had revealed differences in formulaic expression usage, depending on gender of speaker, topic, and discourse setting. To provide comparable normal-control values, a structured interview similar to the contexts utilized for the three patients under investigation was designed and administered to 10 age- and education-matched normal-control subjects. (Example of transcript and analysis is provided in Figure 4.)

The results showed significant differences ($p < 0.05$) between the normal control group (20.1%) and each of the three patients. The subject with subcortical damage showed a frank paucity in that only 11% of words spoken belonged to formulaic expressions; the patient with extensive right-sided damage also yielded a significantly lower proportion at 16.9%, while the subject with transcortical sensory aphasia produced 51.9% in the sample analyzed. Further evaluation of proportions of individual categories in the normal and brain-damaged subjects are currently underway (Sidtis, Canterucci, & Katsnelson, 2008, submitted).

Speech samples freshly obtained from other sources support the notion that formulaic language is amplified in aphasia. In a speech sample provided by Jacqueline Stark,[10] an aphasic subject recovered some speech over a period of

10. Dr. Stark is at the Austrian Academy for Linguistic and Communication Sciences in Vienna, Austria.

Figure 4. Sample of transcript for analysis of speech sample

Utterance	Fixed expressions	Sentence initials	Discourse particles	Pause fillers	Proper names	Numerals
Tell me a little about your family						
Well I was married in xxxxxxx and had xxxx boys over a span of xxxx years.	over a span of		well		xxxx	4, 20, 1941
Uh, like, xxxx's uh in Virginia.				uh, like	xxxx	
He's xxxx.		and				40
And the uh youngest one xxxx 's still at home		and		uh		
And he's twenty.						20
And uh I spent two years in the xxxx during the war				uh	xxxx	2
Uh Grew up in xxxx				uh	xxxx	
Uh and then I went to work for Shell Oil				uh	xxxx	
And I I uh moved xxxx and sa… transferred to xxxx		and		uh	xxxx, xxxx	
And then they dissolved uh the territory so I was without a job		and then		uh		
So uh after four years I went to work for this friend of mine who was an electrical contractor.	this friend of mine	so		uh		4

five treatments from early nonfluency. On inspection, 64% of the speech sample is made up of words in formulaic expressions (below, in italics; novel expressions are underscored).

Table 1. Aphasic speech sample at baseline (Test 1) and after five treatment sessions (Test 5). Formulaic language is in bold italics and novel words and phrases are underscored

Test 1. ***Uh*** **.** ***TV*** **?** ***My*** ***Monday*** ***is*** ***uh*** … ***bank*** ***uh*** **.** ***TV*** **..** ***my*** **..** ***Monday*** ***uh*** ***bank*** **.** ***hm***

Test 5: ***Uh.. uh good morning.. uh.. um.. me uh I want a.. big big ter//uh terevision, alright? Um, big. Alright? And uh.. money? Yes. Fine.. um.. big and. uh … small um.. TV. yes.. uh small um.. Uh.. sky and cricket and.. tennis and.. uh soccer and movies and news and.. alright? Um.. right. Uh.. where? Ah! Alright! Boah! nice! Wow! Big! And small! Ho-ho, Jesus! Uh.. price? What? two thousand.. oh Jesus! hm.. wait. um.. hm hm hm. yes. alright.. um.. I. will uh … I will phone and uh.. uh. woman, yes? And uh um.. wife, yes. Um.. maybe alright.. maybe uh. two thousand? Oh, Jesus. Alright. Uh phone and wait, alright? Uh.. oh, Jesus! Hi! Jane um.. phew.. uh.. what is the matter? Money? Oh, Jesus.. alright.. alright! thank you! see you! Uh salesman.. uh.. money, yes.. fine..***

Another example, showing speech from a German aphasic patient before treatment, was provided by Caterina Breitenstein,[11] who has developed a protocol for intensive speech rehabilitation (Schomacher, Baumgärtner, Winter et al. 2006). Dr. Breitenstein trains subjects in naming and in specific propositional statements useful in activities of daily living. In the initial sample below using the ANELT language evaluation protocol, taken at the baseline condition, nearly all the speech product consists of formulaic language (Blomert, Kean, Koster, and Schokker 1994). A question of interest is this: when propositional speech abilities improve following the intensive training sessions, will formulaic expressions also increase? Studies to answer this question are currently underway.

Table 2. German aphasic speech. Formulaic language is in bold italics and novel words and phrases are underscored (T = Therapist, P = Patient). English translation in italics

T:	Bevor wir anfangen, machen wir einfach mal zwei Übungsbeispiele, ja?
	(Before we begin, let's simply do two practice examples, okay?)
P:	***Ja.***
	(Yes.)
T:	Also, Sie sind bei einem neuen Friseur.
	(Okay. You are at a new hair salon.)
P:	***Ah Gott ja.***
	(Oh heavens yes.)
T:	Und sie sind an der Reihe.
	(And it is your turn.)
P:	***Ja.***
	(Okay.)
T:	Ich bin der Friseur.
	(I am the stylist.)
P:	***Ja.***
	(Okay.)
T:	Was sagen sie zu mir?
	(What do you say to me?)
P:	***Hallo, wie geht's? Danke, gut, tja. ja,*** **und?**
	(Hello, how are you? Thank you, good, okay, yeah, and now?)
T:	Was sagen sie noch?
	(What else do you say?)
P:	***Äh,*** **Haare waschen?Und,** **rot**, ***ja, ja, och,ja.***
	(Uh, wash hair? And, red, yeah, yeah, oh, yeah.)

(Continued)

11. Dr. Breitenstein directs aphasia research in the Neurology Department at the University of Münster, Germany.

Table 2. *Continued*

T:	Noch etwas?
	(Anything else?)
P:	***Nö, äh, ach Gott,*** **und**, ***ein, ehm,*** **und** ***und äh,*** **und** **und**, **Geld**, ***nö, das ist so gut, das ist, das*** **w**..
	(Nope, um, oh God, and, a, um, and and, um, and and, money, nope, that's just fine, that's, that)
T:	Okay, aber es ist richtig. Sie stellen sich vor,
	(Okay, but it is correct to introduce yourself.)
P:	***Ja.***
	(Yes.)
T:	Was wäre wenn,
	(What would it be when…)
P:	***Ja, sehr gut***
	(Okay, very good.)

2.5 Other speech production studies

In one of the few studies comparing formulaic with novel expressions in speech production,[12] propositional and nonpropositional tasks were matched and evaluated in aphasic subjects (Lum & Ellis 1994). Counting was compared to number identification; responsive naming of pictures using cues from formulaic expressions (e.g., "Don't beat around the BUSH") was matched with responsive naming using novel expression cues ("Don't dig behind the BUSH"); and repetition of formulaic expressions was paired with repetition of novel expressions. Better performance on nonpropositional tasks for number production and picture naming but not for phrase repetition was found. This can be explained by the notion that the novel-formulaic distinction pertains to spontaneous processing, and is nullified when a model or template is provided, as in repetition. Van Lancker & Bella (1996) reported similar results in aphasic subjects comparing matched propositional and nonpropositional expressions, with better nonpropositional ability for sentence completion than repetition. Interestingly, careful phonetic analysis of the contrasting repetition tasks did not reveal differences in articulatory skill between the two tasks. This suggested, again, that the mechanisms

12. Studies of comprehension in normal and clinical subjects are more common than production studies; many of these identify the right hemisphere as playing a significant role in various kinds of formulaic language processing, such as idioms and indirect requests (Van Lancker Sidtis 2006b; Weylman, Brownell, Roman & Gardner 1989; Myers 1998).

differentiating propositional and nonpropositional speech modes belong to the spontaneous mode (Van Lancker Sidtis, Ahn, and Yang 2009, in preparation).

Neuroimaging results for speech production related to formulaic expressions of different types have appeared. Early studies using SPECT technology reported bilateral representation of automatic speech tasks (Ryding, Bradvik & Ingvar 1987; Larsen, Skinhøj, and Lassen 1978), but the meaning of these findings is overshadowed by subsequent studies that have reported bilateral brain signals for most language tasks (Van Lancker Sidtis 2006a). Results from imaging studies are not any more consistent for automatic speech than they are for other language tasks.

In one study (Van Lancker, McIntosh & Grafton 2003), five aphasic patients who had suffered a single, unilateral stroke in the perisylvian region were compared to nine right-handed, age- and education-matched normal-control subjects. Tasks were three sets of 90-second activation sessions producing (1) animal names, (2) vocalized syllables, and (3) counting. As expected, behavioral measures differed significantly between normal-controls and patients for generation of animal names, but not for vocalizations or counting. In the normal-control group, greater left frontal activation was identified for naming and nonverbal vocalization, while more RH and basal ganglia areas were identified for counting. For aphasic subjects, naming and nonverbal vocalization were associated with relatively more diffuse and bilateral structures, and counting did not yield a significant brain profile. These results suggested that counting is not strongly lateralized to the left hemisphere as is naming, but caution due to the uncertain meanings of imaging signals must be taken. In an interesting measure of spoken discourse elements, Postman et al. (2007), using a naming task in an functional MRI paradigm, reported right frontal activation during wrong responses, which usually involved an expletive or other formulaic expression. A related study using PET imaging showed a correlation between pause fillers and other marks of dysfluency ("inclusions") and hemispheric side of activation, with more left hemisphere activation in cases of low inclusions (Postman et al. 2006).

Inconsistencies also arise from these studies of formulaic language. A study using PET imaging in normal subjects employed two speech tasks traditionally considered to be automatic: a serial task (months of the year) and a well rehearsed, memorized text (the Pledge of Allegiance) compared to tongue movements and consonant-vowel syllable production (Bookheimer, Zeffiro, Blaxton, Gaillard, and Theodore 2000). Continuous production of the Pledge of Allegiance showed activation in traditional language areas, while reciting the months of the year engaged only limited language areas (Brodmann areas 44 and 22). Tasks did not include counting, which is the automatic speech behavior most frequently preserved in aphasia. In a preliminary report using PET imaging, differences in brain activation patterns for counting compared with storytelling were described (Blank, Scott, and Wise 2001). A later report addressing the same question indicated extensive

bilateral activation for propositional and nonpropositional tasks alike, with no differences in brain sites between speech modes (Blank, Scott, Murphy, Warburton, and Wise 2002). These inconsistencies, again, can be attributed to problems in brain imaging methodology, whereby the meaning of the activation signals is not well understood (Sidtis 2007).

3.1 Summary of neurolinguistic studies: The dual process model

The dual processing model posits that, as is already well known, language is represented in the left hemisphere, and proposes further that formulaic expressions are facilitated by a subcortical-right hemisphere circuit. An implication of subcortical structures in formulaic control arises from several sources: single case studies, speech disorders in subcortical disease, and behavioral functions of subcortical nuclei. Single cases of loss of formulaic language following basal ganglia damage due to stroke have been reviewed above. Swearing and other (taboo) formulaic expressions are hyperactivated in Tourette's disease (Van Lancker & Cummings 1999), which is associated with subcortical dysfunction. In some patients, stimulation of thalamic areas in stereotaxic surgery for treatment of motor disorders elicited recurrent utterances (Petrovici 1980) or "compulsory speech," described as "exclamations… utterances of surprise, fright, or pain," counting, or vocal gestures such as the sound of a shepherd used to collect sheep (Schaltenbrand 1975: 71–3). In one patient, the formulaic expression "thank you" was elicited repeatedly by stimulation of a particular site (Schaltenbrand 1965).[13] Subcortical nuclei (basal ganglia) store and mediate complex motor programs (Marsden 1982), which include vocal motor gestures. Neurological disorders involving these structures could contribute to abnormal diminution or activation of formulaic expressions.

A spotlight is shone on the right hemisphere as playing a key role in formulaic language for several reasons. Propositional speech (grammatical utterances, naming, information-bearing sentences) is disturbed by damage in many areas of the left hemisphere, often preserving production of formulaic expressions. It is reasonable to infer that the right hemisphere supports these expressions. The notion is further supported by the postoperative speech of a normally developing adult whose left hemisphere was removed in medical treatment, presenting discourse markers (well, oh), pause fillers (uh, um), sentence stems (I want), and expletives (God damn it), all produced with normal articulation and prosody, but no other

13. It is difficult to assess the possible formulaic status of other reported utterances elicited during thalamic stimulation, because only English translations are given in the published material.

language (Smith 1966; Van Lancker Sidtis 2004). Furthermore, well-established characteristics of the right hemisphere are compatible with a special role for processing formulaic language (Myers 1998; Van Lancker 1997). Favored are patterns, configurations, and whole complex Gestalts, with more efficient processing of the overall form and content than details or features (Kaplan, Brownell, Jacobs, and Gardner 1990). In communication, contextual meanings are better processed than analytic, linguistic meaning relations. Successful processing of theme and topic as properties of discourse units also requires an intact right hemisphere. An important aspect of many formulaic expressions involves appropriate linguistic and social context. For example "It's a small world" requires a constellation of conditions including chance meeting of acquaintances in an unlikely setting, along with connotations of surprise and so on. This kind of thematic, contextual material has been shown to be preferentially processed by the right hemisphere (Van Lancker 1997; Myers 1998). Finally, establishment of familiarity (personal relevance) appears to be the province of the right hemisphere (Van Lancker 1991).

3.2 Dual process model and schemata

Evidence for a dual process model of language processing comes from several sources, the most compelling of which is neurolinguistic. The implications of these studies are that novel and formulaic language are affected differently by different types of brain damage: left hemisphere damage leads to selective impairment of novel language (with relative preservation of formulaic language), while right hemisphere and/or subcortical damage lead to selective impairment of formulaic language (sparing novel language). Neurological damage can disturb, diminish or enhance behaviors involving formulaic language. Enhancements in formulaic language use are seen in aphasia, Tourette's syndrome, autism, Down's syndrome, and Alzheimer's disease, while diminution is observed in right hemisphere and subcortical disease. It is likely that more such differences will be documented as information about formulaic language is disseminated into clinical practice. Recognition of the important role of formulaic expressions in evaluation and recovery in aphasia and other neurological disorders has barely begun, despite the "automatic speech" tradition extending more than a hundred years into the past.

The notion of two such processing modes has emerged from studies of learning and memory, comparing, for example, procedural and declarative knowledge (Mishkin, Malamut & Bachevalier 1984). Subcortical structures have been associated with "chunking of action repertoires" (Graybiel 1998) or "habit learning" (Knowlton, Mangels, and Squire 1996). These perspectives have been aligned with hierarchical levels of the central nervous system, such that automated motor

gestures are accommodated by subcortical structures, which developed phylogenetically earlier in human evolution (Koestler 1967). Correspondingly, it has been suggested that the origin of human language might be located in initial use of formulaic expressions (see Figure 5 for a whimsical example) (Jaynes 1976; Code 2005; Wray 1998, 2000; Wray & Grace 2007).

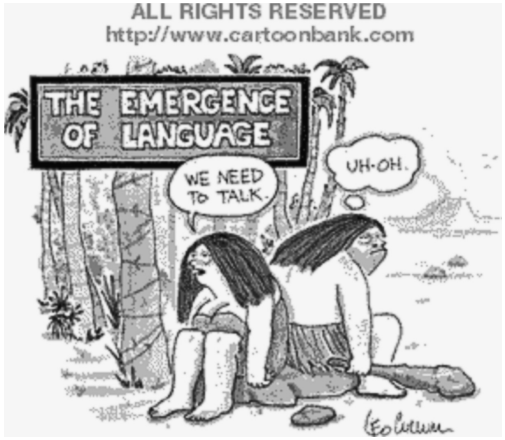


Figure 5. Formulaic expressions may have played a role in human language origins.

Another provocative source that supports the dual-process model arises from developmental language studies, in infants' first and in adult second language acquisition. Researchers in child language document acquisition of holistic "chunks" of speech which evolve into compositional structures (Peters 1983; Lieven 2007; Tomasello 2003). While unitary utterances are utilized by children early on, acquisition of formulaic expressions at adult levels lags behind acquisition of grammatical competence (Kempler, Van Lancker, Marchman, and Bates 1999). This suggests that the two processes, holistic and analytic, perform different roles at different stages of language acquisition, and, further, that different maturational schedules are in play for novel versus formulaic language knowledge. Similarly, in adult second language acquisition, the difficulty posed by formulaic expressions is well known. It is likely that critical periods for native-like acquisition exist for various types of language competences, including for acquisition of formulaic expressions.

Another important source of evidence for the dual-process view is linguistic. As described above, the formulaic phrase has unique properties: it is cohesive and unitary in structure (sometimes with aberrant grammatical form), often nonliteral

or deviant in meaning properties, and usually contains a nuanced meaning that transcends the sum of its (lexical) parts. The canonical form of the expression ("formuleme") is known to native speakers. This is to say that a formulaic expression functions differently in form, meaning and use from a matched literal, novel, or propositional expression (Lounsbury 1963). "It broke the ice," for example, as a formula, differs regarding meaning representation, exploitation of lexical items, status in language memory,[14] and range of possible usages, when compared to the exact same sequence of words as a novel expression.

3.3 Comparison of formulaic expressions with schemata

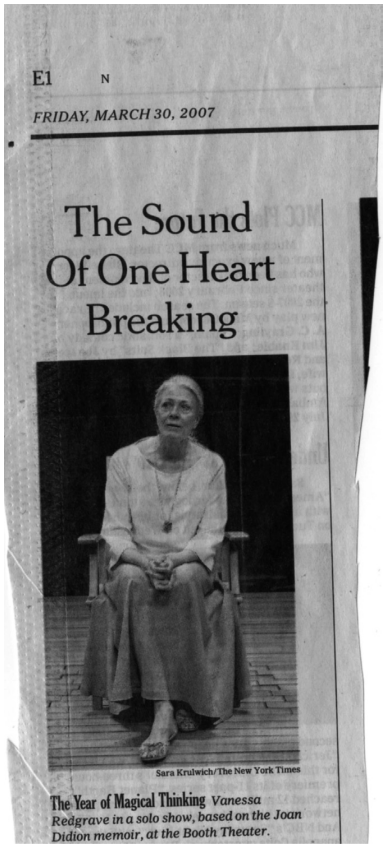
A primary property of formulaic expressions is their cohesion or unitary structure. This has led to their characterization as lexical units. However, considerable flexibility has also been described (Sprenger 2003), such that morphemes and words can be inserted and grammatical rules applied under various circumstances, as in "She had him totally eating right out of her hand," or, to tersely describe a grisly death scene, "The bucket was certainly kicked here." The formulaic unit can be alluded to by mentioning only a portion, as in "I wouldn't want to be counting chickens…"[15] Changes can be applied to formulemes in humor, word games, or other kinds of language play, so long as their canonical shape remains recognizable. In Figure 6, two words are replaced in part of an utterance that alludes to a well-known Zen koan or philosophical riddle, "What is the sound of one hand clapping?" and thereby drawing on the nuances of mystery and intensity inherent in that saying.

One approach to modeling the structural properties of novel and formulaic language is to view expression-types as occurring at two extremes, from fixed, in which the underlying formuleme is known, and novel, where word choices are dependent on grammatical constraints only. An intermediate type of expression is the schema (Lyons 1968: 177–8).[16]

14. The unique states of formulaic expressions in memory storage accounts for the "idiom effect" seen in word association studies (Clark 1970).

15. Also observed in field studies: "Besides the small world thing,…" and "All that stuff you see…is just frosting on that basic cake."

16. The internet site www.languagehat.com describes a similar phenomenon using the term "snowclones," which include formulaic expressions and schemata used in journalistic writing and speeches. See also Wray 2000.


E1 N

FRIDAY, MARCH 30, 2007

The Sound Of One Heart Breaking

Sara Krulwich/The New York Times

The Year of Magical Thinking *Vanessa Redgrave in a solo show, based on the Joan Didion memoir, at the Booth Theater.*

Figure 6. Example of lexical replacement in formulaic expression ("What is the sound of one hand clapping?") (The New York Times).

A schema is a fixed form with one or more free open slots. Schemata carry the characteristics of formulaic expressions in having a basic canonical form (with distinctive intonation contour and voice quality), utilizing specialized meanings, conveying nuances, and being known, but they have an additional versatility (See Appendix II for a sample list). Examples are "That was a _____ and a half," and "If you had my _____, you'd be _____, too." A preliminary collection of 209 schemata reveals a range of word-count lengths from 1–19 words, with a mean utterance length of 4.74 words (Figure 7), and a mean of 1.25 open slots.

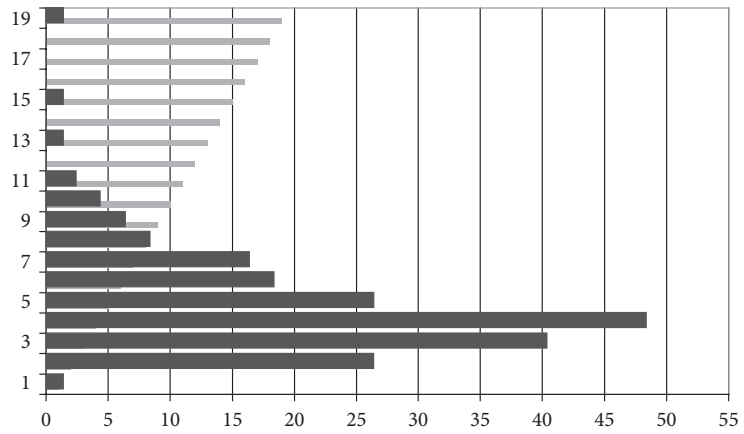


Figure 7. Schemata word count per expression is represented on the Y axis, and frequency on the X axis.

In schemata, two processing modes, novel and formulaic, are creatively interactive. A known unitary form, a formulaic expression, allows specific flexibility in accommodating novel expression. Here "the best of both worlds" is in play. Schemata vividly illustrate the dual process in linguistic performance, in which two distinct modes, analytic and holistic, coexist in continuous interplay. It is the claim of the dual-process model that two different modes of language processing can be seen in child language acquisition, differential effects of neurological damage, psycholinguistic studies, and everyday language use. These concepts have relevance for theoretical and practical models of language behavior.

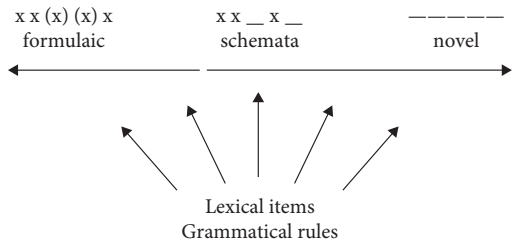


Figure 8. Structural status of formulaic, schematic, & novel utterances: Morphophonemic & movement rules, lexical insertion.

Appendices

Appendix I. Some categories of formulaic language with German counterparts.

1. Idioms: light at the end of the tunnel.
   Wendungen: Licht am Ende des Tunnels.
2. Proverbs: Rome wasn't built in a day.
   Sprichwörter: Rom wurde auch nicht in einem Tag erbaut.
3. Slang: Awesome, cool
   Umgangsprache: Geil, cool.
4. Conventional expressions (various types): What luck, in the meantime, more or less, as I was saying
   Formeln: Glück im Unglück, sozusagen, mehr oder minder, Wie gesagt,
5. Speech formulas: How are you? See you later.
   Floskeln: Guck mal an. Tschüss.
6. Indirect requests: It's getting late. Isn't it kind of warm in here?
   Indirekte Forderungen: Es wird spät. Es ist hier so warm oder bin ich das?
7. Expletives: Good heavens, jumpin' Jimminy.
   Schimpfwörter: Donnerwetter, Um Gottes Willen.
8. Sentence stems: I'd like you to meet…, I want
   Satzanfänge: Ich möchte..,
9. Memorized expressions: Prayers, rhymes, songs
   Auswendig gelernte Ausdrücke: Gebete, Lieder
10. Serial speech: numbers, alphabet
    Seriensprache: Zählen, Alphabet
11. Pause fillers: well, ya know
    Pausenfüller: also, wissen Sie
12. Familiar proper nouns: George W. Bush.
    Eigenname: Gerhard Schröder
13. Discourse markers: uh, um
    Füllworter, Füllsel, Verzögerungswörter: äh

Appendix II. Selected schemata

1. _____ and counting
2. _____ to end all _____
3. A _____ does not a _____ make.
4. A ___ without_____ is like a _____ without _____
5. A _____'s _____ (word repeated)
6. A walking_____
7. A whole nother _____

8. Do I look like a _____ ?
9. Down with _____
10. Goodbye _____, hello _____
11. Have enough _____ there?
12. He is too _____ by half
13. How _____ is that?
14. I (he) eat(s) and breathe(s) _____
15. I can do _____ with one hand tied behind my back.
16. I eat _____ for breakfast.
17. I know _____ like the back of my hand.
18. I may not know anything about _____, but I know what I like.
19. I wouldn't be caught dead _____
20. I wouldn't give you _____ for his _____
21. If you had his/my _____, you'd be _____(-ing) too.
22. I'll give you a _____
23. I'm all _____ed out.
24. I'm not a big _____ person
25. It was (a)_____ from hell.
26. It's not just about (the) _____; it's about (the) _____
27. It's nothing if not _____
28. It's(he's, she's) a little too _____ by half
29. Leave the _____ at home
30. Make like a _____ and _____.
31. mother of all _____
32. Move over, _____.
33. My middle name is _____
34. None of this _____ business
35. now that's a _____
36. Shut up and _____
37. So you think you can _____
38. Some of my best friends are _____
39. That was a _____ and a half
40. That was voted the most _____
41. The_____ are taking over.
42. Those wacky _____
43. To think I was once (a) _____
44. Using the _____ word
45. Wadda I look like, a _____ ?
46. We know _____ when we hear (see) it
47. What if _____ is what it's all about?
48. What part of _____ don't you understand?
49. What's up with _____
50. When _____ is not enough
51. You (I) must have been absent when they handed out the _____
52. You call that a _____?

53. You can say hello to_____, goodbye to _____
54. You can take (your)_____ and shove it.
55. You can take the ___ out of the ___, but you can't take the ___ out of the ___.
56. You've got to love the _____
57. You've seen one _____, you've seen them all.

References

Alajouanine, Theophile. 1956. Verbal realization in aphasia. *Brain* 79: 1–28.

Basso, Anna. 2003. *Aphasia and its therapy*. Oxford: OUP.

Bay, E. 1964a. Aphasia and conceptual thinking. In *Problems of dynamic neurology,* L. Halpern (Ed.), Jerusalem: Hebrew University Hadassah Medical School.

Benson, D. Frank. 1979. *Aphasia, alexia, and agraphia*. New York NY: Churchill Livingston.

Berthier, Marcelo L. 1999. *Transcortical aphasias*. Hove: Psychology Press.

Blank, S. Catrin, Sophie K. Scott & Richard J.S. Wise, R. 2001. Neural systems involved in propositional and non-propositional speech. Abstract 11192; Human Brain Mapping Conference, Brighton, UK.

Blank, S. Catrin, Sophie K. Scott, Kevin Murphy, Elizabeth Warburton & Richard J.S. Wise. 2002. Speech production: Wernicke, Broca and beyond. *Brain* 125: 1829–1838.

Blanken, Gerhard & Marini, V. 1997. Where do lexical speech automatisms come from? *Journal of Neurolinguistics* 10: 19–31.

Blanken, Gerhard. 1991. The functional basis of speech automatisms (recurring utterances). *Aphasiology* 5: 103–127.

Blomert, Leo, M.-L. Kean, C. Koster & Schokker, J. 1994. Amsterdam-Nijmegen Everyday Language Test: Construction, reliability and validity. *Aphasiology* 8: 381–407.

Bookheimer, Susan. Y., Zeffiro, T.A., Blaxton, T.A., Gaillard, P.W., & Theodore, W.H. 2000. Activation of language cortex with automatic speech tasks. *Neurology* 55: 1151–7.

Chung, Kevin K.H., Chris Code & Martin J. Ball. 2004. Lexical and non-lexical speech automatisms in aphasic Cantonese speakers. *Journal of Multilingual Communication Disorders* 2: 32–42.

Clark, Herbert H., 1970. Word associations and linguistic theory. In *New horizons in linguistics*, Lyons, J. (Ed.), 271–286. Baltimore MD: Penguin Books.

Code, C. 2005. First in, last out? The evolution of aphasic lexical speech automatisms to agrammatism and the evolution of human communication. *Interaction Studies* 6: 311–334.

Code, Chris. 1982. Neurolinguistic analysis of recurrent utterance in aphasia. *Cortex* 18: 141–152.

Critchley, MacDonald. 1970. *Aphasiology and other aspects of language*. London: Edward Arnold.

Dieguez, Sebastian & Julien Bogousslavsky. 2007. Baudelaire's aphasia: From poetry to cursing. In *Neurological disorders in famous artists*, Part 2 [Frontiers of Neurology and Neuroscience 22], J. Bogousslavsky & M.G. Hennerici (Eds), 121–149. Basel: Karger.

Espir, Michael L.E. & Clifford Rose, F. 1970. *The basic neurology of speech*. Oxford: Blackwell Scientific Publications.

Fillmore, Charles. 1979. On fluency. In *Individual differences in language ability and language behavior,* C.J. Fillmore, D. Kempler & W.S-Y Wang (Eds), 85–102. London: Academic Press.

Gloning, I., Gloning, K. & Hoff, H. 1963. Aphasia – A clinical syndrome. In *Problems of dynamic neurology*, L. Halpern (Ed.), Jerusalem: Hebrew University Hadassah Medical School.
Goldstein, Kurt. 1948. *Language and language disturbances.* New York NY: Grune and Stratton.
Goodglass, Harold & Edith Kaplan. 1972. *The assessment of aphasia and related disorders*. Philadelphia PA: Lea and Febiger.
Goodglass, Harold & Mayer, J. 1958. Agrammatism in aphasia. *Journal of Speech and Hearing Disorders* 23: 99–111.
Graves, Roger & Theodore Landis. 1985. Hemispheric control of speech expression in aphasia. *Archives of Neurology* 42: 249–251.
Graybiel, Ann M. 1998. The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory* 70: 119–136.
Head, Henry. 1926. *Aphasia and kindred disorders of speech*. Cambridge: The University Press.
Hughlings Jackson, John. 1874 [1932]. On the nature of the duality of the brain. In *Selected Writings of John Hughlings Jackson*, Vol. 2, J. Taylor, (Ed.), 129–145. London: Hodder and Stoughton.
Jaynes, Julian. 1976. *The origin of consciousness in the breakdown of the bicameral mind*. Boston MA : Houghton Mifflin.
Kaplan, J.A., Brownell, H.H., Jacobs, J.R. & Gardner, H. 1990. The effects of right hemisphere damage on the pragmatic interpretation of conversational remarks. *Brain and Language* 38: 122–134.
Kempler, Daniel, Diana Van Lancker, Virgina Marchman & Elizabeth Bates. 1999. Idiom comprehension in children and adults with unilateral brain damage. *Developmental Neuropsychology* 15.3: 327–349.
Knowlton, Barbara, Jennifer A. Mangels & Larry R. Squire. 1996. A neostriatal habit learning system in humans. *Science* 273: 1399–1402.
Koestler, Arthur. 1967. *The ghost in the machine*. Chicago IL: Henry Regnery Company.
Larsen, B., Skinhøj, E. & Lassen, H.A. 1978. Variations in regional cortical blood flow in the right and left hemispheres during automatic speech. *Brain* 10: 193–200.
Lieven, Elena. 2007. Producing multiword utterances. In *Constructions in acquisition,* B. Kelly & E. Clark (Eds) Stanford CA: CSLI.
Lounsbury, Frank.G. 1963. Linguistics and psychology. In *Psychology: Study of a science,* 553–582. New York NY: McGraw-Hill.
Lum, C.C. & Ellis, A.W. 1994. Is 'nonpropositional' speech preserved in aphasia? *Brain and Language* 46: 368–391.
Luria, Alexander R. 1964. Factors and forms of aphasia. In *Disorders of Language: CIBA Symposium,* A.V.S. DeReuck & M. O'Connor (Eds), London: J. and A. Churchill.
Luria, Alexander R. 1966. *Higher cortical functions in man*. New York NY: Basic Books.
Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge: CUP.
Marsden, C.D. 1982. The mysterious motor function of the basal ganglia: The Robert Wartenberg lecture. *Neurology* 32: 514–539
Mishkin, M., Malamut, B. & Bachevalier, J. 1984. Memories and habits: Two neural systems. In *Neurobiology of Learning and Memory,* G. Lynch, J.L. McGaugh & N.M. Weinberger (Eds), 65–67. New York NY: The Guilford Press.
Myers, Penelope. 1998. *Right hemisphere damage: Disorders of communication and cognition*. San Diego CA: Singular Press.
Peters, Ann M. 1983. *The units of language acquisition.* Cambridge: CUP.

Petrovici, J.-N. 1980. Speech disturbances following stereotaxic surgery in ventrolateral thalamus. *Neurosurgical Review* 3(3): 189–195.
Pick, Arnold. 1973. *Aphasia*. Springfield IL; Charles C. Thomas. English translation by J. Brown, from *Handbuch der Normalen und Pathologischen Physiologie*, 1931, 15: 1416–1524.
Pinker, Stephen. 1995. *The language instinct*. New York NY: HarperCollins.
Postman, Whitney A., Birn, R., Pursley, R., Butman, J., McArdle, J., Xu, J., Solomon J. & Braun, A. 2007. When right is wrong: An fMRI study of overt naming in patients with aphasia. 2007 Annual Meeting of the Cognitive Neuroscience Society, New York, NY.
Postman, Whitney A., Solomon, J., Maisog, J., Chapman, S.B., Tuttle, S., Christian, M.R., Milosky, L. & Braun, A. 2006. Deconstructing discourse: A PET study of narrative production. 19th Annual CUNY Conference on Human Sentence Processing. New York NY.
Ryding, E., Bradvik, B. & Ingvar, D. 1987. Changes of regional cerebral blood flow measured simultaneously in the right and left hemisphere during automatic speech and humming. *Brain* 110: 1345–1358.
Schaltenbrand, George. 1965. The effects of stereotactic electrical stimulation in the depth of the brain. *Brain* 88: 835–840.
Schaltenbrand, George. 1975. The effects on speech and language of stereotactical stimulation in thalamus and corpus callosum. *Brain and Language* 2: 70–77.
Schomacher, M., Baumgärtner, A. Winter, B. Lohmann, H. Dobel, C. Wedler, K. Abel, S. Knecht, S. & C. Breitenstein. 2006. Erste Ergebnisse zur Effektivität eines intensiven und hochfrequent repetitiven Nennen- und Konversationstrainings bei Aphasie. *Forum Logopädie* 4 (20): 22–28
Sidtis, John J. 2007. Some problems for representations of brain organization based on "activation." *Brain and Language* 102 (2): 130–140.
Sidtis, D., Canterucci, G. & Katsnelson, D. 2008. Effects of neurological damage on production of formulaic damage, submitted.
Smith, A. 1966. Speech and other functions after left (dominant) hemispherectomy. *Journal of Neurology, Neurosurgery and Psychiatry* 29: 467–471.
Speedie, Lynn J., Wertman, E., T'air, J. & Heilman, K.M. 1993. Disruption of automatic speech following a right basal ganglia lesion. *Neurology* 43: 1768–1774.
Sprenger, Simone A. 2003. Fixed expressions and the production of idioms. *MPI Series in Psycholinguistics*. 21. Wageningen: Ponsen and Looijen.
Tomasello, Michael. 2003. *Constructing a language*. Cambridge MA: Harvard University Press.
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): 1–11.
Van Lancker Sidtis, Diana & Whitney A. Postman. 2006. Formulaic expressions in spontaneous speech of left- and right-hemisphere damaged subjects. *Aphasiology* 20(5): 411–426.
Van Lancker Sidtis, Diana & Gail Rallon. 2004. Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Language and Communication* 24: 207–240.
Van Lancker Sidtis, Diana. 2001. Preserved formulaic expressions in a case of transcortical sensory aphasia compared to incidence in normal everyday speech. *Brain and Language* 79 (1): 38–41.
Van Lancker Sidtis, Diana. 2004. When novel sentences spoken or heard for the first time in the history of the universe are not enough: Toward a dual-process model of language. *International Journal of Language and Communication Disorders* 39(1): 1–44.

Van Lancker Sidtis, Diana. 2006a. Has neuroimaging solved the problems of neurolinguistics? *Brain and Language* 98: 276–290.
Van Lancker Sidtis, Diana. 2006b. Where in the brain is nonliteral language? *Metaphor and Symbol* 21(4): 213–244.
Van Lancker Sidtis, Diana, Ji Sook Ahn & Sueng-Yun Yang. 2008. Novel and formulaic language following unilateral stroke: production and comprehension studies. In preparation.
Van Lancker, Diana. 1988. Nonpropositional speech: Neurolinguistic studies. *In Progress in the psychology of language*, Vol. III, A.E. Ellis (Ed.), 49–118. Hillsdale NJ: Lawrence Erlbaum Associates.
Van Lancker, Diana & R. Bella. 1996. The relative roles of repetition and sentence completion tasks in revealing superior speech abilities in patients with nonfluent aphasia. *Journal of the International Neuropsychological Society* 2: 6.
Van Lancker, Diana & Jeffrey Cummings. 1999. Expletives: Neurolinguistic and neurobehavioral perspectives on swearing. *Brain Research Reviews* 31: 83–104.
Van Lancker, Diana. 1973. Language lateralization and grammars. In *Studies in syntax and semantics*, Vol. II, J. Kimball (Ed.), 197–204. New York NY: Academic Press.
Van Lancker, Diana. 1975. Heterogeneity in Language and Speech: Neurolinguistic Studies. *Working Papers in Phonetics* 29. Los Angeles CA: UCLA.
Van Lancker, Diana. 1991. Personal relevance and the human right hemisphere. *Brain and Cognition* 17: 64–92.
Van Lancker, Diana. 1993. Nonpropositional speech in aphasia. In *Linguistic disorders and pathologies: An international handbook,* G. Blanken, J. Dittmann, J. Grimm, J.C. Marshall, C-W. Wallesch (Eds), 215–225. Berlin: Walter de Gruyter.
Van Lancker, Diana, Bogen, J.E. & Canter, G.J. 1983. A case report of pathological rule-governed syllable intrusion. *Brain and Language* 20: 12–20.
Van Lancker, Diana, J. Canter & Terbeek, D. 1981. Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech and Hearing Research* 24: 330–335.
Van Lancker, Diana, R. McIntosh & Grafton, R. 2003. PET activation studies comparing two speech tasks widely used in surgical mapping. *Brain and Language* 85: 245–261.
Van Lancker-Sidtis, Diana 2003. Auditory recognition of idioms by first and second speakers of English: It takes one to know one. *Applied Psycholinguistics* 24: 45–57.
Weylman, S.T., Brownell, H.H. Roman, M. & Gardner, H. 1989. Appreciation of indirect request by left- and right-brain damaged patients: The effects of verbal context and conventionality of wording. *Brain and Language* 36: 580–591.
Wilder, W.I. 1959. Director. *Some like it hot*, Film, starring Jack Lemmon, Tony Curtis, and Marilyn Monroe.
Wilder, W.I. & Diamond, A.L. 1959. *Some like it hot*. Screenplay reprinted in S. Thomas (Ed.). 1990. *Best American screenplays* 2, 1st edn), 80–146. New York NY: Crown Publishers.
Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.
Wray, Alison & George W. Grace. 2007. The consequences of talking to strangers: Evolutional corollaries of socio-cultural influences on linguistic form. *Lingua* 117: 543–578.
Wray, Alison. 1998. Protolanguage as a holistic system for social interaction. *Language and 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*, C. Knight, J.R. Hurford & M. Studdert-Kennedy (Eds), 285–302. Cambridge: CUP.

PART II

Psychological reality

The psycholinguistic reality of collocation and semantic prosody (2)

Affective priming

Nick C. Ellis & Eric Frey
English Language Institute, University of Michigan

Abstract

We investigate the psycholinguistic reality of the corpus linguistic phenomena of collocation and semantic prosody. Ellis, Frey & Jalkanen (in press) used lexical decision tasks to demonstrate that word recognition processes were sensitive to collocation, but not semantic prosody. The current research used an affective priming task to investigate whether semantic prosody affected later stages of semantic processing. Verbs' semantic prosody correlated with conceptual evaluations of their pleasantness. Verbs positive or negative in semantic prosody caused significant affective priming, effects that were independent of conceptual evaluation. We conclude that people acquire through language usage implicit knowledge of the types of word with which verbs collocate, and this can facilitate subsequent semantic processing of material which accords with these usage norms.

1. Introduction

Corpus linguistics has clearly demonstrated that natural language makes considerable use of recurrent patterns of words and larger constructions. Lexical context is crucial to knowledge of word meaning and grammatical role. One type of pattern is *collocation*, described by Firth (1957) as the characterization of a word from the words that typically co-occur with it. Sinclair (1991: 100), summarized the results of corpus investigations of such distributional regularities in the *Principle of Idiom:* "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments," and suggested that for normal texts, the first mode of analysis to be applied is the idiom principle, as most of text is interpretable by this principle. Kjellmer (1987: 140) reached a similar conclusion: "In all kinds of texts, collocations are indispensable elements with which our utterances are very largely made". Erman & Warren (2000) estimate that about half of fluent native text is constructed according to the idiom principle. Comparisons of written and spoken corpora suggest that collocations are even more frequent in spoken language (Biber, Johansson, Leech, Conrad & Finegan 1999; Brazil 1995; Leech 2000).

Collocations are patterns of preferred co-occurrence of particular words, like *blazing row* and *heated dispute* (but not *heated row* or *blazing dispute*). Other patterns, deriving from generalization across collocations, are more abstract. *Semantic prosody* refers to the general tendency of certain words to co-occur with either negative or positive expressions, "the consistent aura of meaning with which a form is imbued by its collocates" (Louw 1993: 157). A famous example, by Sinclair, is *set in*, which has a negative prosody: *rot* is a prime exemplar for what is going to set in. *Cause* (something causes an accident/catastrophe/other negative event), *commit* (suicide, crime, offence), and *happen* (things go along smoothly, then 'something happens', shit happens) similarly have a negative semantic prosody. These patterns come from usage – there are no defining aspects of the meaning of *cause*, *commit*, or *happen* which entails that they will take negative rather than positive objects. Hoey (2005) refers to such generalizations when a word or word sequence is associated in the mind of a language user with a semantic set or class as *semantic association*.

Corpus linguistic and cognitive linguistic analyses of the phenomena of collocation, formulaic language, semantic prosody, and other aspects of phraseology in language *texts* demonstrate how lexis, grammar, meaning and usage are inseparable (Ellis 2008a,b; Granger & Meunier 2008; Hunston & Francis 1996; Robinson & Ellis 2008; Sinclair 1991, 2004). Such observations have naturally provoked inferences about language *users* and about the cognitive processes of meaning, speech production and comprehension. The statement of the *Principle of Idiom* is a good example, others include:

1. Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*… (Firth 1957: 196)
2. In the store of familiar collocations there are expressions for a wide range of familiar concepts and speech acts, and the speaker is able to retrieve these as wholes or as automatic chains from the long-term memory; by doing this he minimizes the amount of clause-internal encoding work to be done and frees himself to attend to other tasks in talk-exchange, including the planning of larger units of discourse… (Pawley & Syder 1983: 192).
3. … for a great deal of the time anyway, language production consists of piecing together the ready-made units appropriate for a particular situation and … comprehension relies on knowing which of these patterns to predict in these situations. (Nattinger 1980: 341).

(4) Every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word. If one of the effects of the initial priming is that regular word sequences are constructed, these are also in turn primed… The(se) are claims about the way language is acquired and used in specific situations. (Hoey 2005: 13)

(5) Corpus-based analysis can throw light on the nature and extent of collocational bonding between words… In addition, data of the kind considered here can reveal something of the cognitive processes which lie behind language learning and use, and which enable us to become fluent language users, and it is these insights which can be among the most satisfying of all. (Kennedy 2003: 485)

But these statements overstep the data. While there is no denying that texts have been produced by language users, and thus must somehow reflect their thinking, corpus analyses say nothing about the cognitive loci of sensitivity of language users to these patterns of co-occurrence. The analysis of whether word recognition and lexical access, semantic activation, and the processes of production of speech and writing are sensitive to collocations and the more abstract schemata potentially derivable from them is an empirical matter, one that falls into the domain of investigation of psycholinguistics.

Psycholinguistic research broadly confirms language users' sensitivity to various distributional aspects of orthographic, phonological, morphological and syntactic form (Ellis 2002a, 2002b, 2008a, 2008b): There are effects of bigram frequency in visual word identification and of phonotactic knowledge in speech segmentation, effects of spelling-to-sound correspondences in reading, and cohort effects in spoken word recognition. There are effects of neighbors and the proportion of friends (items which share surface pattern cue and have the same interpretation) to enemies (items which share surface pattern but have different interpretations) in

reading and spelling, morphology, and spoken word recognition. At higher levels, it can be shown that language comprehension is determined by the listeners' considerable knowledge of the statistical behavior of the lexical items in their language. In comprehending language, people make use of their knowledge of the relative frequencies with which individual verbs appear in different tenses, in active vs. passive structures, and in intransitive vs. transitive structures, the typical kinds of subjects and objects that a verb takes, and many other such facts, and thus they perceive the most probable syntactic and semantic analyses of a new utterance on the basis of frequencies of previously perceived utterance analyses (Seidenberg 1997). In production too, language users tend to generate the most probable utterance for a given meaning on the basis of frequencies of utterance-representations. Thus it has been argued that "Psycholinguistics is the testament of rational language processing and the usage model" (Ellis 2005, 2006).

Nevertheless, psycholinguistic research also identifies a wide variety of largely separable processes of language cognition (Altman 1997; Gernsbacher 1994), and it demonstrates that these are *differentially* affected by factors such as type and token frequency, phonological, orthographic, morphosyntactic, grammatical and pragmatic consistency of pattern, cohort density and consistency, word class, imageability, age of acquisition, etc. (Harley 1995; Levelt 1989). Our research program therefore investigates the degree to which various broad neighborhoods of language processing are affected by these patterns of collocation and semantic prosody identified by corpus linguists. We use the processing divisions illustrated in Figure 1 – word recognition and lexical access, semantic processing, and speech production – and determine whether these are separately sensitive (1) to particular patterns of collocation, and (2) to the abstract generalizations of semantic prosody, in order to determine the psycholinguistic reality of these textual phenomena. This is a large enterprise and we have therefore attacked it piecemeal.

The first stage of our work (Ellis, Frey & Jalkanen in press) investigated the effects of these phenomena upon lexical access. We found that processing in a lexical decision task, where two letter strings were presented simultaneously and the participant had to decide whether both were words or not (Meyer & Schvaneveldt 1971), was clearly sensitive to patterns of usage of booster/maximizer-adjective and verb argument collocations. Native speakers were quicker to decide that *blameless* was a word when it followed a frequent collocate like *entirely,* or *mauled* following *badly,* than when the same pool of words was re-sorted as controls which contained the same words combined randomly, thus removing the sequential patterning of English collocational usage (e.g., *badly blameless, entirely mauled*) while nevertheless maintaining sense and grammaticality. They were similarly faster to decide that *maturity* was a word when it followed a frequent verb collocate like *attain* than they were when it followed a non-collocate like *cause*. Given that the lexical

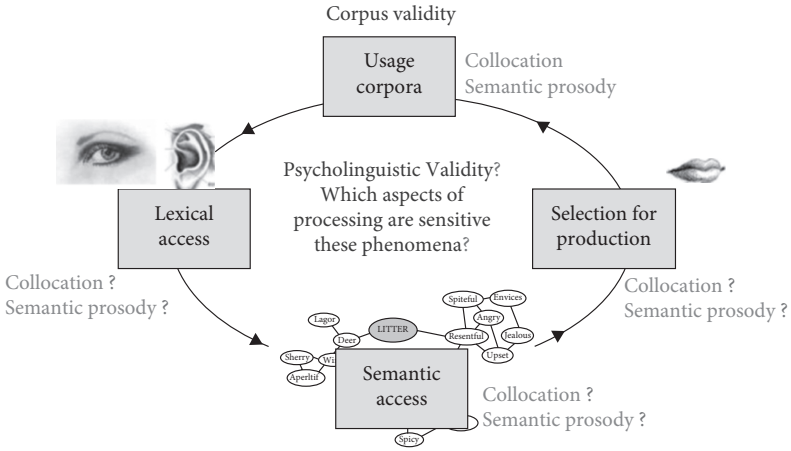


Figure 1. The bounds of investigation: To what extent are these different psycholinguistic processes sensitive to the separate corpus-valid phenomena of collocation and semantic prosody?

decision task minimally requires word recognition and access to the lexicon, we concluded that these processes are tuned by experience of particular collocations in usage, so that higher frequency collocations are more readily perceived than lower-frequency ones. The language recognition system tallies the co-occurrence of these particular words in usage (Ellis 2002a) and so tunes itself accordingly to preferentially process them as collocations on future encounters. But this research also showed that the same paradigm which so readily showed sensitivity to particular collocations failed to demonstrate generalization – people were no faster at judging that *good* was a word when preceded by a verb like *attain* that did not specifically collocate with it, but which nevertheless was strongly of a matching semantic prosody. Thus we concluded that there were no top-down generalizations upon the level of processing required for lexical access.

The current experiment therefore extends the investigation of semantic prosody deeper into the system (Figure 1) by determining the degree to which it might affect semantic access.

Our measure of semantic prosody was grounded in the work of Kjellmer (2005) whose analyses of patterns of collocation of English verbs in the BNC allowed him to identify twenty verbs that were strongly negative in their semantic prosody (e.g., *cause*: something causes an accident/catastrophe/other negative outcome) and twenty strongly positive verbs (e.g., *achieve*: one achieves objectives/goals/success/other positive outcomes). We took these verbs as candidate stimuli and

then operationalized various corpus statistics measuring direction and strength of semantic prosody, as described in the method section below, in order to determine the degree to which fluency of semantic access is affected by prosodic valence.

Our investigation of semantic processing was based on the affective priming paradigm, a psycholinguistic technique for investigating implicit positive or negative attributions. Fazio, Sanbonmatsu, Powell, and Kardes (1986) reasoned that a priming effect similar to that found with lexical decision should also be apparent for automatic evaluative attitudinal semantics. Presentation of an attitude object (any object – *spider*, *alcohol*, *The President*, or whatever) as a prime should activate any associated evaluations and, hence, facilitate a related judgment. The paradigm that Fazio et al. (1986) developed, and that has been commonly employed since, involved participants' performance on an adjective connotation task. The target word presented on each trial is an evaluative adjective (any adjective, for example, *pleasant*, *frightening*, *corrupt, incompetent*) and participants are instructed to indicate whether the word is positive or negative as quickly as possible. The focus of these experiments was on the latency with which this judgment is made and, in particular, the extent to which it is facilitated by the presentation of an attitude object as a prime. In three experiments, Fazio et al. (1986) found evidence of automatic attitude activation. Responding was faster on trials for which the participants' evaluations of the primed attitude objects were congruent with the connotation of the targets than on trials for which they were incongruent. For example, if the attitude object *pain* is evaluated negatively by an individual, then presentation of *pain* as the prime automatically activates the negative evaluation. If a subsequently presented target adjective is also negative (e.g., *disturbing*), then the individual is able to indicate the connotation of the target adjective relatively quickly, more so than if a positive adjective (e.g., *appealing*) serves as the target word. Subsequent research (De Houwer & Hermans 2001; De Houwer, Hermans, Rothermund & Wentura 2002; Fazio 2001; Hermans, De Houwer & Eelen 1994) shows this to be a robust phenomenon, although the size of the effect does vary as a result of stimulus exposure times and their stimulus onset asynchrony (SOA), stimulus type (words, pictures, etc.), and the nature of the response (evaluation, naming, etc.). For evaluative categorization, brief SOA's reveal stronger priming effects (Hermans, De Houwer & Eelen 2001). The subsequent lore of affective priming research using evaluative responses for word stimuli has it that it is best to use SOAs of 150 or 200 ms., i.e., to present primes for 150 or 200 ms. and have the target immediately following prime offset (without an inter-stimulus interval), to use an external response box since keyboards can introduce a lot of error in the latencies, and to register response latencies as well as error data and to analyze them as a composite measure because effects are often distributed over these two dependent variables.

The current experiment thus used an affective priming task to measure the speed and accuracy with which participants rate a target word as generally positive (pleasant) or negative (unpleasant), and to see if reaction time and accuracy were affected by the degree to which a prime matched the target in semantic prosody. A composite measure of these two dependent variables (AccSpeed) was made by summing the standarized scores for accuracy and speed, with positive values of the composite AccSpeed measure reflecting good performance and negative values reflecting bad performance. We predicted that target words with a positive valence would be processed faster and more accurately after verbs with positive semantic prosody than those with a negative semantic prosody, and conversely, that words with a negative valence would be processed faster and more accurately after verbs with a negative semantic prosody than verbs with a positive semantic prosody.

A related question of interest, that of Firth (1957) quoted above, concerned the dissociable contributions of conceptual and syntagmatic knowledge to semantics. Propositional meaning, perceptual reference, and syntagmatic usage provide three different sources of word meaning. (1) *Propositionally*, a dog is, by definition, a canine, any of various fissiped mammals with nonretractile claws. (2) *Referentially*, the word *dog* automatically awakens perceptual memories, sights, touches, smells, and these imagery associations affect our understanding. Words with high imageability are represented not only propositionally but also in an imagery code, as "sensory images awakened" (James 1890). "Concrete terms such as *house* readily evoke both images and words as associative (meaning) reactions, whereas abstract words such as *truth* more readily arouse only verbal associations. The meaning of the latter is primarily intraverbal." (Paivio 1971: 85). Propositional meanings and imagery associations have been shown to be dissociable and additive sources of meaning and memory in a wide range of cognitive psychological (Ellis 1991; Paivio 1990), and brain imaging studies (Pulvermüller 1999), as well as in neuropsychological dysfunction. For example, Warrington (1975; 1981) describes three cases of visual object agnosia where there was impairment in knowledge of pictorial representations of objects from visual presentation *and* from memory, where knowledge of subordinate categories was more vulnerable than superordinate categories ("to refer to the often-quoted example of the *canary*, these patients could correctly categorize it as living, animal, and bird [the attributes of these superordinates still being known] but could not reliably classify it as yellow, small and pet", [1975: 655]; other examples included *bucket* being defined as 'container', but on further questioning no details of its size, weight or function, and *pigeon* -> 'I know it is a bird but not which one'). Given that these symptoms could neither be accounted for by intellectual impairment, sensory or perceptual deficit, or expressive language disorder, Warrington argues that there are two functionally distinct modality-specific meaning systems, i.e., a particular concept, say *canary*, would be

represented in two semantic memory hierarchies, the one primarily visual and the other primarily verbal. These cases' cerebral lesions result in their loss of the former while preserving the latter – visually imageable words have become abstract. (3) *Syntagmatically*, the word *dog* also awakens associations with words experienced as its common collocates in language usage, with meaning deriving from the company it keeps with *walk*, *leash*, *vet*, and even *hot*, and *tired* (Firth 1957; Hoey 2005).

Usually, since language describes the world, these three sources of meaning converge, which is why corpus analytic techniques like Latent Semantic Analysis put words into the same meaning space as do more conceptual analyses (Landauer & Dumais 1997). But syntagmatic and paradigmatic evidence do not always align. As already mentioned, *lack* is negatively evaluated yet has a positive semantic prosody in that its collocates are all positive (*lack resources, lack money, lack experience*), while *arouse* and *cure* are positive in their semantics but of a negative semantic prosody. There follows a variety of interesting psycholinguistic questions relating to the effects on their processing of nice words like *cure* falling into the bad company of *cancer*, *disease*, *ills* and the like.

In this particular study, we hoped to exploit these dissociations to investigate whether affective priming is a conceptual phenomenon arising from matching meanings, or a syntagmatic one stemming from experience of collocations. Thus, we also gathered participants' explicit ratings of pleasantness for the verbs in order to determine whether corpus-derived semantic prosody measures or subjective evaluations of the emotional valence were better predictors of affective priming.

In summary, our specific goal was to determine whether fluent language users have implicit knowledge of semantic prosody that is automatically brought to bear as a top-down facilitative influence in the semantic processing of language input which accords to these usage norms.

2. Experiment: The effects of a verb's semantic prosody on semantic processing

2.1 Method

2.1.1 *Participants*

The experiment involved 15 adult volunteers (9 male, 6 female) recruited from the student population of the University of Michigan-Ann Arbor. They were native speakers of English aged around 20 years (M = 20.9, SD = 1.7). They were paid $10 for their participation.

2.1.2 *Materials*

Verbs judged to have strong positive and negative semantic prosody were selected as follows. Kjellmer (2005) analyzed 20 positive and 20 negative semantically prosodic verbs and described methods of determining their degree by assessing their most frequent collocates and the relative numbers of these that were positive or negative. After he kindly sent us a draft list of these verbs, we developed these operationalizations further. Each usage of these verbs was determined in the British National Corpus (BNC) using Davies' (2007) interface (http://corpus.byu.edu/bnc/): (1) All collocates following the verb within 3 words were extracted. We recorded the frequencies of the verb, the frequencies of the words with which it collocated, and the frequencies of the particular collocations themselves. We ordered the latter by decreasing frequency. (2) For all collocations with token frequency >= 2, or the top 500 most frequent of these if more than that, two independent raters judged each collocate for whether they thought it was positive, neutral, or negative. These raters, one of whom is the second author of this study, were undergraduates studying topics in psychology, linguistics, and anthropology. Interpretation of words out of context is variable; this indeed is the central theme of the Idiom Principle and of constructional/phraseological approaches to language, thus there was some variability in these judgments. Nevertheless, the two raters showed enough accord to warrant continuation: the inter-rater agreement was 79% for the positive items, and 85% for the negative items. For each verb we then summed the number of positive, negative, and neutral collocates and computed a variety of indices of prosodic valence and strength, including the total number of collocate types of the verb's valence, the percentage of overall collocate types that were of its valence, and its ratio of positive to negative collocate types. Pooling these various indices, we selected ten strongly positively semantically prosodic of the original verb set: *restore, attain, live, achieve, guarantee, advise, grant, gain, regain, lend*, and ten strongly negative: *wreak, inflict, contract, battle, commit, provoke, wage, suffer, cause, cure*. These and their collocation analyses are shown in Table 1.

Each of these twenty verbs were then combined with various other words as stimuli for an affective priming task based on the paradigm of Fazio et al. (1986) and De Houwer et al. (2002) in which participants were briefly presented a prime followed by a target noun, which they were asked to rate as either positive or negative.

Some of the paired items involved specific collocates of the verbs. These included matched pairs (made with the two most common collocates of the polarity of the particular prime, e.g., *attain-goals, attain-maturity, cause-problems, cause-damage*) and two mismatched pairs (made with the two most common collocates of a prime of opposite polarity, e.g., *attain-problems, attain-damage,*

Table 1. Determination of semantic prosody

Prime	Frequency (per million words) as verb	Frequency (per million words) all tokens	Semantic prosody valence	Total n collocates of that valence	% of all collocates of that valence	Ratio +/– collocates
attain	452	452	+	41	37	13.7
cause	5738	12876	–	568	57	0.1
lack	1009	9871	+	121	41	11.0
cure	521	1472	–	55	72	0.0
gain	3663	5137	+	316	32	5.1
suffer	3421	3421	–	400	58	0.1
guarantee	1435	3911	+	108	30	8.3
fight	3871	6706	–	194	30	0.4
grant	1294	7594	+	106	32	3.3
provoke	588	588	–	74	51	0.1
restore	1648	1648	+	197	26	7.0
encounter	667	1670	–	12	29	0.2
lend	1254	1254	+	42	24	6.0
ease	1078	3020	–	120	49	0.1
achieve	6715	6715	+	321	32	6.2
contract	505	11882	–	26	30	0.3
secure	2773	4548	+	250	32	6.4
commit	1339	1341	–	78	44	0.1
emphasize	654	654	+	57	24	4.1
arouse	310	310	–	26	41	0.3

cause-goals, cause-maturity). This generated a total of 80 prime-target pairings, with 40 'positive' responses and 40 'negative' responses (see Table 2).

To assess semantic prosody/association rather than specific collocation, each verb was also paired with four generalization items of positive valence (*good, benefit*, *virtue,* and the emoticon ☺, generating, e.g., the polarity matching *attain-good, attain-benefit, attain-virtue, attain-*☺ and mismatching *cause-good, cause-benefit, cause-virtue,* and *cause-*☺), four generalization items of negative valence (*bad, harm*, *evil*, and ☹ generating, e.g., the polarity mismatching *attain-bad, attain-harm*, etc. and matching *cause-bad, cause-harm,* etc.). This created a total of 160 prime-target pairings, with 80 'positive' responses and 80 'negative' responses.

In all, the experiment thus involved 240 prime-target pairings. During the task, presentation consisted of one prime-target pairing at a time, and trials were randomized for each participant to avoid potential order effects.

Table 2. Prime-target pairings with the top collocates

	Matched collocates		Mis-Matched collocates	
Prime	Target 1	Target 2	Target 1	Target 2
attain	goals	maturity	problems	damage
cause	problems	damage	goals	maturity
lack	confidence	resources	problems	disease
cure	problems	disease	confidence	resources
gain	access	understanding	loss	damage
suffer	loss	damage	access	understanding
guarantee	success	safety	war	battle
fight	war	battle	success	safety
grant	permission	relief	crisis	violence
provoke	crisis	violence	permission	relief
restore	confidence	pride	problems	difficulties
encounter	problems	difficulties	confidence	pride
lend	hand	support	pain	burden
ease	pain	burden	hand	support
achieve	success	growth	cancer	disease
contract	cancer	disease	success	growth
secure	knowledge	access	suicide	offence
commit	suicide	offence	knowledge	access
emphasize	importance	value	suspicion	controversy
arouse	suspicion	controversy	importance	value

The present paper concerns semantic prosody and so we need to restrict our analyses to participants' performance *on generalization items only*. It is important that we are not looking at effects of specific collocation. Therefore we went back to the BNC and checked for any particular occurrences of collocation between our verbs and the generalization items *good, benefit*, *virtue, bad, harm*, and *evil* in a 3 subsequent word window. Whenever such collocations were evident (e.g., 36 occurrences of *gain* + *benefit*), we removed this pair from the analysis. There were 37 collocation types so identified. All the analyses in this paper are therefore restricted to the 123 prime-target trials which involve novel verb-object pairings that are not found in the BNC.

2.2 Procedure

The task was programmed in E-prime (Schneider, Eschman & Zuccolotto 2002) running under Windows XP on standard desktop PCs. SuperLab response boxes

were used as the input device, allowing participants' reaction times to be recorded with millisecond accuracy.

Upon arrival at the lab, participants were briefed about the nature of the experiment (to investigate people's knowledge of the ways words combine in English phrases) and asked to sign a consent form. They were instructed as follows: "On each trial you will see a word on the computer screen followed by a colored target word. Your task is to judge whether this colored word is positive or negative. You will be asked to do this as quickly and accurately as possible by pressing the relevant key on the keyboard". There followed an initial practice session of 12 verb-object/prime-target pairings consisting of non-study words. This gave participants a chance to get used to the requirements of the task. The experiment instructions were shown again before the main session of 240 prime-target trials began. The task measured how fast and accurately participants judged a target word (e.g., *goals, confidence*) to be generally positive (pleasant) or negative (unpleasant), and assessed whether they did this faster when it was primed by a verb of the matching valence of semantic prosody (e.g., *attain-goals*, *lack-confidence, cause-evil*) than by a mismatching one (e.g., *attain-problems*, *lack-disease, cause-benefit*). On each trial, the verb prime was presented for 200 ms, followed immediately by a target word to be rated as either positive (1) or negative (2) on the SR box. They were given a maximum of 2000 ms to input an answer. A brief pause followed each response with an on-screen message reading "Press SPACE BAR to Continue" so that they could rest between trials as they felt appropriate. The trial sequence is illustrated in Figure 2. The program recorded individual reaction times (in milliseconds) and accuracy of response. In order to combine accuracy and reaction time into one measure (AccSpeed), we standardized the reaction time and accuracy data. Standardized variables (z-scores) have a mean of zero and a standard deviation of 1. High values of the standardized accuracy measure reflect good performance. High values of RT reflect bad performance, and so the z-scores on RT were multiplied by –1 to turn them into a speed measure. A composite measure was then made by summing the z-scores for accuracy and the z-score for speed. Thus positive values of the composite AccSpeed measure reflect good performance and negative values reflect bad performance.

In a supplemental task after the main experiment, participants were asked to rate the 20 verb primes on a nine point scale of pleasantness from most positive (+4) to most negative (–4). The verbs were presented individually mid-screen in randomized order and the participants were given as long as they wished to consider and rate them. These explicit ratings indexed the degree to which respondents assessed the verbs to be emotionally positive or negative in their conceptual meaning.

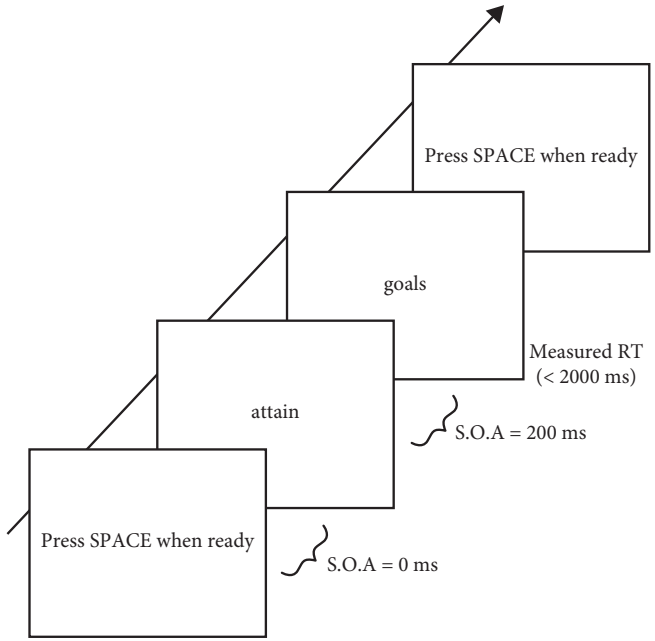


Figure 2. Sequence of presentation in affective priming task.

Our four specific questions, which direct the sections of the Results section, were as follows:

Question 1. To what extent are semantic prosody and conceptual meaning associated, and to what extent can they be dissociated? In our discussion of lexical semantics, above, we identified the separate contributions of syntagmatics (collocation and semantic prosody) and reference. In our experiments we separately measured these – the syntagmatics in the corpus analyses, the referential aspects in respondents' explicit evaluations of verb pleasantness. Regression analyses can thus be used to determine the degree to which these two measures are associated.

Question 2. Are there measurable effects of the semantic prosody of verbs upon speed and accuracy of semantic processing of subsequent words in an affective priming task? Regression analyses can investigate the association between semantic prosody and reaction time and, separately, between semantic prosody and AccSpeed. Effects of congruence between prime and target in the affective priming task should show themselves as high values on AccSpeed when the negative generalization items (*bad, harm*, *evil*, and ☹) are primed by more negative

semantic prosody verbs, and decreasing AccSpeed when these negative targets are primed by verbs of increasing positive semantic prosody. Equally, there should be low values of AccSpeed when positive generalization items (*good, benefit*, *virtue,* and ☺) are primed by verbs of more negative semantic prosody, and increasing AccSpeed values the more positive the semantic prosody of the verbs. These predictions are illustrated in the top panel of Figure 5. The predicted slope is negative for the negative generalization items and positive for the positive generalization items. The critical test of semantic prosody effects, therefore, is whether these two regression lines are of opposite sign and differ significantly from each other.

Question 3. Are there measurable effects of the conceptual evaluations of verbs upon speed and accuracy of semantic processing of subsequent words in an affective priming task? As for Question 2, the test of congruence is whether there is greater AccSpeed when more negatively evaluated verbs precede negative generalization items and when more positively evaluated verbs precede positive generalization items. The critical tests is again whether there is a significant slope difference between these two graphs.

Question 4. Are there independent affective priming effects of semantic prosody and conceptual meaning? This question involves the determination of whether semantic prosody explains additional variance in AccSpeed beyond conceptual evaluation, and vice versa. Hierarchical regression analysis is the appropriate technique here. Thus, for example, step 1 might involve the regression of AccSpeed upon conceptual evaluation and then, with this relationship statistically controlled, step 2 could test whether there is significant extra prediction of AccSpeed if semantic prosody is then entered into the equation.

2.3 Results

2.3.1 *The relationship between semantic prosody and conceptual meaning*

In order to assess the relationship between semantic prosody, as operationalized in our corpus analyses, and subjective explicit evaluations of conceptual meaning, we averaged the participants' ratings of the verbs' pleasantness in the supplemental tasks and plotted these against two measures of semantic prosody. The top panel of Figure 3 shows the relationship with the absolute number of positive or negative collocates where there is a strong positive relationship between conceptual meaning and semantic prosody ($\beta = 0.57$, $p < .001$, $R^2 = 0.32$). The bottom panel shows the relationship with the percentage of total collocations which were positive or negative; again the relationship is positive and significant ($\beta = 0.29$, $p < .001$, $R^2 = 0.08$). It is clear that, for the present sample of 20 verbs at least, semantic prosody and conceptual meaning are positively associated, although

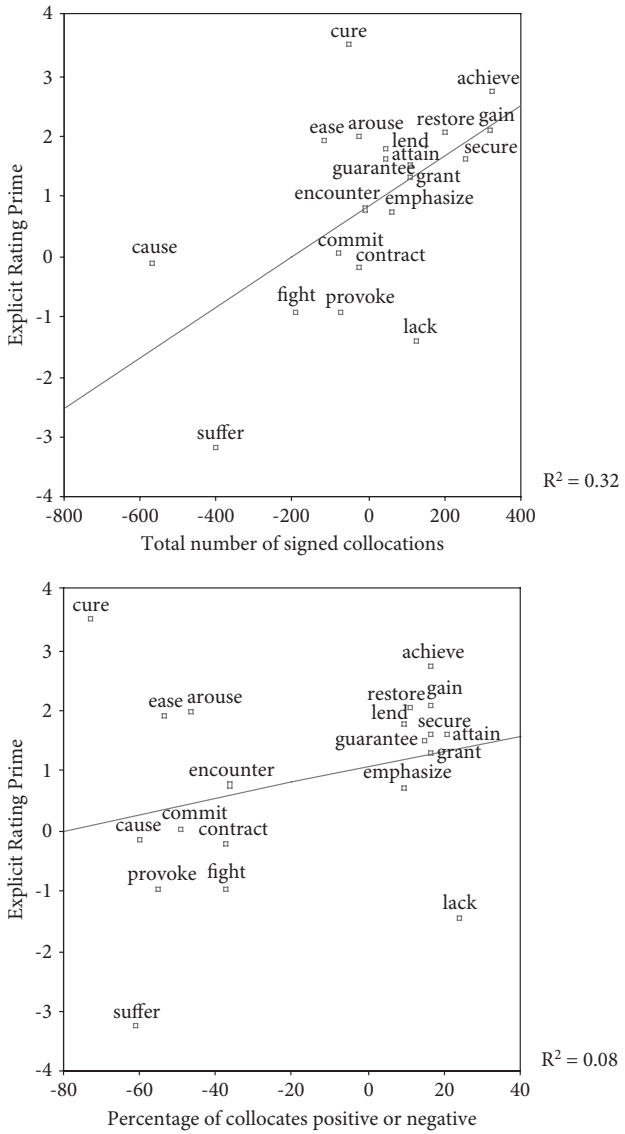


Figure 3. The relationship between participants' explicit ratings of the pleasantness of the verbs and their semantic prosody as defined (top) as the number of positive (negative) collocates in the BNC, or (bottom) the percent of collocates which were positive (negative).

there are odd exceptions to this rule, particularly *cure*, a positively evaluated word which is of negative semantic prosody, and *lack*, a negatively evaluated word that is of strong positive semantic prosody.

2.3.2 *The effect of semantic prosody on affective priming*

We operationalized affective priming in two ways, firstly in terms of effects upon response time, and secondly, since effects can be distributed across both latency and accuracy, upon their composite measure AccSpeed.

Response time. For each verb prime we calculated the mean reaction time for all positive generalization items minus the mean speed rating for all negative generalization items. Thus, greater priming of positive targets results in a more negative value, greater priming of negative targets in a more positive value, and little, if any priming benefit results in a mean value close to zero.

The top panel of Figure 4 shows the association with the absolute number of positive or negative collocates, where there is a negative relationship between semantic prosody and RT Mean difference ($\beta = -0.36$, $R^2 = 0.13$, n.s.). The bottom panel shows the association with the percentage of total collocations which were positive or negative; again the relationship is negative ($\beta = -0.27$, $R^2 = 0.07$, n.s.).

In both analyses, as the number positive collocates increases, so does the priming advantage for positive targets over negative targets; equally, as the number of negative collocates increases, so there is a priming advantage of negative targets over positive ones. However, despite explaining 13% and 7% of the variance in RT difference respectively, neither of these regressions reaches significance.

Response time and accuracy composite. As explained above for Question 2 and illustrated above the graphs in Figure 5, the test of congruence is whether there is greater AccSpeed when verbs of more negative prosody precede negative generalization items and when verbs of more positive semantic prosody precede positive generalization items.

The graphs in Figure 5 follow the predicted patterns of affective priming for semantic generalizations, for both positive and negative generalized items. For the negative generalization items (left panel), a linear regression shows a negative correlation between the composite accuracy-speed score, AccSpeed, and increasingly positive semantic prosody ($\beta = -0.22$, $p = 0.07$), explaining roughly 5% of the variance. For the positive generalized items, linear regression shows a positive correlation between strength of positive semantic prosody and AccSpeed ($\beta = .26$, $p = .06$), explaining about 7% of the variance.

The major test of our predictions is whether the slopes of the two regression lines, that for the negative and positive generalization items, differ significantly

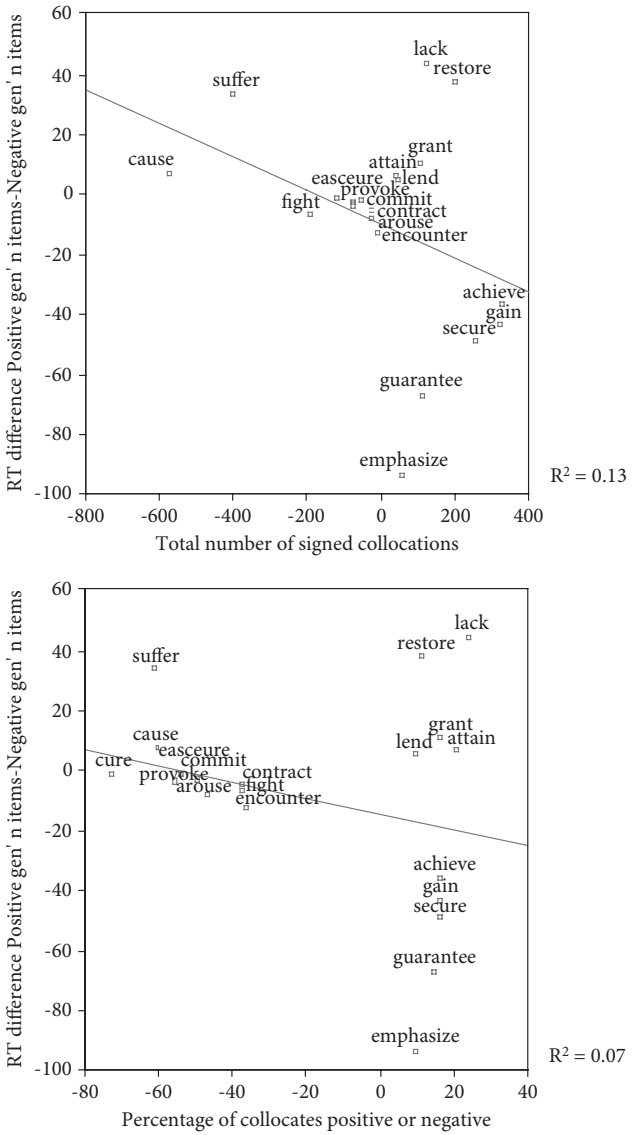


Figure 4. The mean difference of reaction times (ms.) between positive generalization items and negative generalization items plotted as a function of two different indices of semantic prosody: total number of signed collocates (top), percentage of signed collocates (bottom). Points labeled by prime.

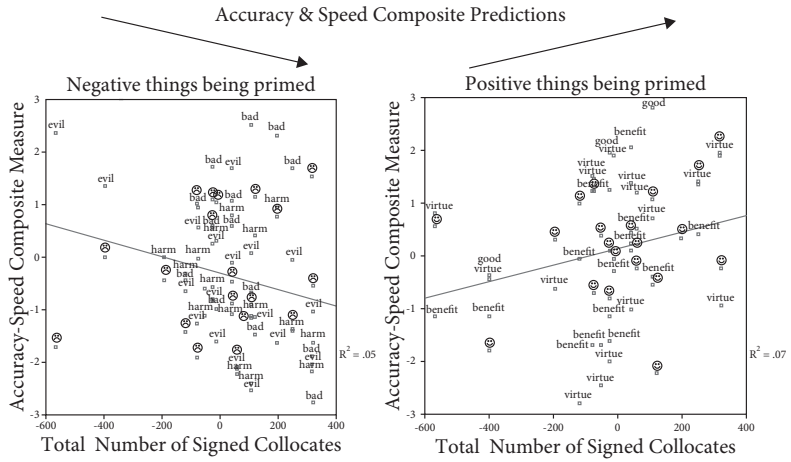


Figure 5. Accuracy and Speed in evaluating "Negative" generalization targets (left panel) and "Positive" generalization targets (right panel) as a function of the semantic prosody of the prime measured as total number of signed collocates.

from each other. We tested the difference between these two correlations following the procedure outlined in Howell (1982: 197–198) and this was indeed the case ($z = 2.67$, $p < .01$).

These data thus demonstrate affective priming results where the affective valence of the prime (in this case determined by semantic prosody of the verb) is, in the interpretation of Fazio et al. (1986), automatically awakened upon its presentation. Remember that in this task participants did not have to rate the prime, indeed they were not oriented to the primes at all and there was no systematic relationship between primes and target since the design had them match and mismatch in valence 50 % of the time following a random, unpredictable, schedule. Nevertheless, on trials where prime and target matched in valence, accuracy and speed was superior to that when they mismatched.

These results suggest that the affective value of a verb prime is automatically and quickly (it is at least initiated within 200 ms.) activated, thus to facilitate the semantic evaluation of subsequent words. Given that the affective value of the verb primes here is defined corpus linguistically in terms of the percentage of overall collocation objects of the verb that were positive (or negative), we conclude that semantic prosody has psychological reality in that the semantic prosody of a verb is automatically accessed and its spreading activation automatically affects the processing of subsequent material.

2.3.3 *The effects of conceptual meaning upon affective priming*

While there is no denying the effects observed in Section 2.3.2, there is still the possibility that they are attributable to a confounding source. In the introduction we discussed the tendency for syntagmatic and conceptual relations to be positively associated. This should come as no surprise since language evolved to describe the world. Thus, nice words tend to go with nice words, just as the nice things they relate to tend to co occur. As we showed in Section 2.3.1, this applied to our sample too, with the correlation between corpus-derived measures of semantic prosody and participants' conscious evaluations of whether words are conceptually positive or negative being $r = 0.57$, $p < .001$.

Could it be, therefore, that it is the conceptual meaning of the primes that is driving affective priming rather than their semantic prosody?

In order to determine this, we ran the same analyses as in Section 2.3.2, but with the participants' evaluations of the affective valence of the verbs as the predictor variable rather than their semantic prosody. The results are shown in Figure 6.

The graphs in Figure 6 also follow the predicted patterns of affective priming for semantic generalizations, for both positive and negative generalized items. For the negative generalization items (left panel), a linear regression shows a negative correlation between the composite accuracy-speed score, AccSpeed,

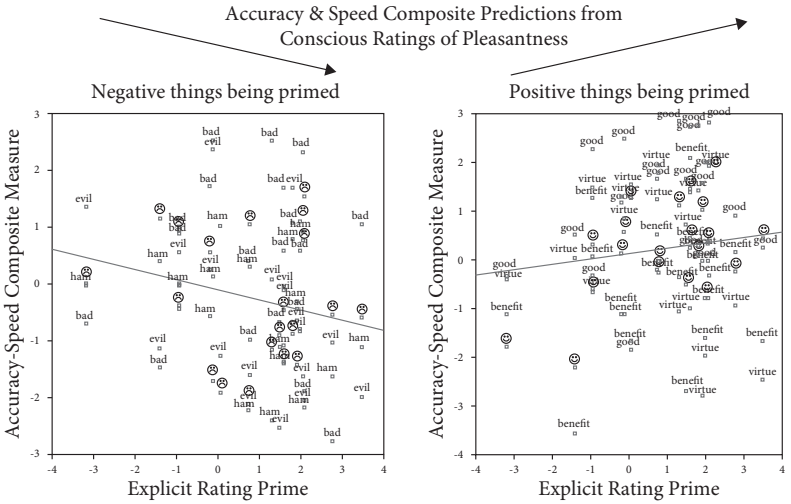


Figure 6. Accuracy and Speed in evaluting "Negative" generalization targets (left panel) and "Positive" generalization targets (right panel) as a function of the Participant's consciously considered conceptual evaluation of the prime as measured in the subsidiary task.

and increasingly positive semantic prosody (β = –0.21, p = 0.09), explaining roughly 4% of the variance. For the positive generalized items, linear regression shows a marginally positive correlation between strength of positive semantic prosody and AccSpeed (β = .01, ns), although this fails to differ significantly from a slope of zero.

As with the semantic prosody results, here too the two correlations, that for the negative generalization items and that for the positive ones, differ significantly from each other, albeit only just so (z = 1.67, p < .05 one tailed).

2.3.4 *Direct comparisons of conceptual meaning and semantic priming*

Combined analyses. It is possible to align the combined accuracy and speed data for the positive and negative generalization items, simply by multiplying those for the negative generalization items by –1. Then the data for all 123 generalization trials where there were no actual collocations between the prime and target in the BNC can be analyzed at once. When we do this we see that semantic prosody (signed N +/ –) correlates with aligned AccSpeed r = 0.25, p < .01), and that explicit rating correlates with aligned AccSpeed r = 0.16, p = .08).

We used hierarchical stepwise regression to determine whether semantic prosody or explicit rating were independently associated with AccSpeed. When semantic prosody was entered first in a multiple regression equation predicting aligned AccSpeed as the dependent variable it was a significant predictor (β = 0.25, p < 0.01) and stepwise regression failed to enter explicit rating at a second stage. However, when explicit rating was entered first (β = 0.16, p = 0.08), explaining only .025 of the variance in AccSpeed, stepwise regression entered semantic prosody at a second stage (β = 0.25, p = 0.01) with this second model explaining .062 of the overall variance in AccSpeed. This additional variance explained by semantic prosody on top of that provided by explicit rating was significant at p < .05.

3. Conclusions

The primary aim of this experiment was to investigate the degree to which native language users are sensitive to semantic prosody in their language processing. In the affective priming task Section 2.3.2, the accuracy and speed with which participants judged target words to be semantically positive or negative was consistently superior when these were primed by verbs of a matching rather than mismatching valence of semantic prosody. In the combined analyses of Section 2.3.4, semantic prosody correlated with aligned AccSpeed r = 0.25, p < .01. The standard interpretation of affective priming (Fazio et al 1986) is that the affective value of the prime

is implicitly and automatically activated, thus to facilitate the semantic evaluation of the subsequent target. Given that the affective value of the verb primes in this experiment were defined corpus linguistically in terms of the percentage of overall collocation objects of the verb that were positive (or negative), we must conclude that the corpus-derived concept has psychological reality in that the semantic prosody of a verb is automatically, implicitly, and quickly (it is at least initiated within 200 ms.) accessed and its spreading activation automatically gives top-down support in the semantic processing of subsequent material that accords with usage norms.

We also observed that the 'aura of meaning' imbued upon words by their collocates is usually in accord with their conceptual meaning – people usually evaluate words of positive semantic prosody as pleasant, and words of negative semantic prosody to be unpleasant. In our small sample in Section 2.3.1 the correlation was r = 0.57, p < .001. It is a challenge therefore to disentangle the contributions of these two sources of meaning and this important confound raises the general need for caution in the interpretation of any effects of corpus derived measures of semantic prosody.

Given the inseparability of grammar and lexis, and that of grammar and meaning, as corpus linguistic, cognitive linguistic, and phraseological analyses have so pervasively demonstrated (e.g., Conklin & Schmitt 2007; Ellis 2008a; Ellis, Simpson-Vlach & Maynard 2008), it should come as no surprise that a word's semantic prosody is entangled with its conceptual meaning. Nevertheless, there are good theoretical motivations for trying to disentangle their effects at different levels of psycholinguistic processing. When we directly assessed the effects of conceptual meaning upon accuracy and speed in the affective priming tasks we obtained a correlation between participants' explicit ratings of verb pleasantness and aligned AccSpeed r = 0.16, p = .08, only marginally significant. Furthermore, the stepwise regressions of Section 2.3.4 demonstrate that while semantic prosody has significant effects upon AccSpeed above those of explicit rating, the reverse is not true. Comparing these two causal variables in this experiment, therefore, we must conclude that semantic prosody has both a numerical and statistically significant edge over conceptual meaning in its effects upon the semantic processing of subsequent words in this affective priming task.

Ellis, Frey & Jalkanen (in press) found that lexical decision was sensitive to patterns of collocation, and thus concluded that processes of word recognition and lexical access are tuned by experience of combinations of particular words in usage, so that higher probability collocations are more readily perceived than lower-frequency ones. The language recognition system tallies (Ellis 2002a) the co-occurrence of these particular words in usage and tunes itself accordingly to preferentially process them as collocations on future encounters. Thus the corpus

linguistic phenomenon of collocation is psycholinguistically real, evidencing itself in processing as early as word recognition.

But lexical decision was not sensitive to semantic prosody – Ellis, Frey and Jalkanen (in press) could identify no such top-down effects upon processes of word recognition. However, the current experiment gives credence to the psycholinguistic reality of this corpus linguistic phenomenon too: there are effects of semantic prosody, albeit later in processing, at semantic access.

Such psycholinguistic validation of phraseological analyses has important consequences for our understanding of language as a dynamic system (Bybee & Hopper 2001; de Bot, Lowie & Verspoor 2007; Ellis 2007, 2008a; Ellis & Larsen-Freeman 2006, 2009; Larsen-Freeman 1997; MacWhinney 1999) wherein there are rich interactive effects of language use, language processing, language learning, and language structure. Usage shapes our construction of mental grammars, mental lexicons and meaning (Goldberg 2006; Hoey 2005; Langacker 2000; Robinson & Ellis 2008; Tomasello 2003). Language users have an extensive implicit knowledge of particular language sequences. The mental lexicon (Elman 2004) and the mental grammar (Spivey 2006) are entirely dynamic and contextualized, with processing ever sensitive to the sequential dependencies experienced in usage (Christiansen & Chater 2001; Ellis 2002a, 2009; Seidenberg & MacDonald 1999).

Acknowledgements

We are extremely grateful to Göran Kjellmer for making the results of his corpus analyses available for use in this study, to Dirk Hermans for detailed advice on the affective priming paradigm, and to Gregory Garretson, the ELI research sharing group, Roberta Corrigan as Editor, and attendees at the International Conference on Exploring the Lexis-Grammar Interface, Hanover, Germany, 5–7 October, 2006 and the Symposium on Formulaic Language, University of Wisconsin-Milwaukee, March 18–21, 2007 for comments on this work.

References

Altman, Gerry T. 1997. *The ascent of Babel*. Oxford: OUP.

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

Brazil, David. 1995. *A grammar of speech*. Oxford: OUP.

Bybee, Joan L. & Paul Hopper (Eds), 2001. *Frequency and the emergence of linguistic structure* [Typological Studies in Language 45]. Amsterdam: John Benjamins.

Christiansen, Morten H. & Nick Chater (Eds), 2001. *Connectionist psycholinguistics*. Westport CT: Ablex.

Conklin, Kathy & Norbert Schmitt. 2007. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 28: 1–18.

Davies, Mark. 2007. View: Variation in English words and phrases. http://view.byu.edu

de Bot, Kees, Wander Lowie & Marjolijn Verspoor. 2007. A dynamic systems theory to second language acquisition. *Bilingualism: Language and Cognition* 10: 7–21.

De Houwer, Jan & Dirk Hermans. 2001. Editorial: Automatic affective processing. *Cognition and Emotion* 15: 113–114.

De Houwer, Jan, Dirk Hermans, Klaus Rothermund & Dirk Wentura. 2002. Affective priming of semantic categorisation responses. *Cognition and Emotion* 16: 643–666.

Ellis, Nick C. 1991. In verbal memory the eyes see vividly, but ears only faintly hear, fingers barely feel and the nose doesn't know: Meaning and the links between the verbal system and modalities of perception and imagery. In *Imagery and cognition,* R.H. Logie & M. Denis (Eds), 313–329. Edinburgh: Plenum Press.

Ellis, Nick C. 2002a. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2): 143–188.

Ellis, Nick C. 2002b. Reflections on frequency effects in language processing. *Studies in Second Language Acquisition* 24(2): 297–339.

Ellis, Nick C. 2005. Usage-based and form-focused SLA: The implicit and explicit learning of constructions. In *Language in the context of use: Usage-based approaches to language and language learning*, Andrea Tyler, Yiyoung Kim, Mari Takada & Diana Marinova (Eds), 1–28. Berlin: Mouton de Gruyter.

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1): 1–24.

Ellis, Nick C. 2007. Dynamic systems and SLA: The wood and the trees. *Bilingualism: Language & Cognition* 10: 23–25.

Ellis, Nick C. 2008a. Phraseology: The periphery and the heart of language. In *Phraseology in foreign language learning and teaching*, F. Meunier & S. Granger (Eds), 1–13. Amsterdam: John Benjamins.

Ellis, Nick C. 2008b. The dynamics of language use, language change, and first and second language acquisition. *Modern Language Journal* 92: 2.

Ellis, Nick C. In press. Words and their usage: Commentary on the Special Issue on the Bilingual Mental Lexicon, *The Mental Lexicon*, 3: 3.

Ellis, Nick C., Eric Frey & Isaac Jalkanen. In press. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In *Exploring the lexis-grammar Interface*, U. Römer & R. Schulze, (Eds), Amsterdam: John Benjamins.

Ellis, Nick C. & Diane Larsen-Freeman. 2006. Language emergence: Implications for applied linguistics. *Applied Linguistics* 27(4) whole issue.

Ellis, Nick C. Diane Larsen-Freeman (Eds), 2009. *Language as a complex adaptive system*. Special Issue, *Language Learning*, 59: Supplement 1.

Ellis, Nick C., Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly.* Special Issue on Psycholinguistics and TESOL. 42: 3, 375–396.

Elman, Jeff L. 2004. An alternative view of the mental lexicon. *Trends in Cognitive Science* 8: 301–306.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20: 29–62.
Fazio, Russell H. 2001. On the automatic activation of associated evaluations: An overview. *Cognition and Emotion* 15: 115–141.
Fazio, Russell H, David M. Sanbonmatsu, Martha C. Powell & Frank R. Kardes. 1986. On the automatic activation of attitudes. *Journal of Personality and Social Psychology* 50: 229–238.
Firth, John R. 1957. *Papers in linguistics: 1934–1951*. London: OUP.
Gernsbacher, Morton A. 1994. *A handbook of psycholinguistics.* San Diego CA: Academic Press.
Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: OUP.
Granger, Sylviane & Fanny Meunier (Eds), 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.
Harley, Trevor A. 1995. *The psychology of language: From data to theory.* Hove: Taylor & Francis.
Hermans, Dirk, Jan De Houwer & Paul Eelen. 1994. The affective priming effect: Automatic activation of evaluative information in memory. *Cognition and Emotion* 8: 515–533.
Hermans, Dirk, Jan De Houwer & Paul Eelen. 2001. A time course analysis of the affective priming effect. *Cognition and Emotion* 15: 143–165.
Hoey, Michael P. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.
Howell, David C. 1982. *Statistical methods for psychology.* Boston MA: Wadsworth.
Hunston, Susan & Gill Francis. 1996. *Pattern grammar: A corpus driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
James, William. 1890. *The principles of psychology,* Vol. 1. New York NY: Holt.
Kennedy, Graeme. 2003. Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly* 37: 477–486.
Kjellmer, Göran. 1987. Aspects of English collocations. In *Corpus linguistics and beyond,* W. Meijs (Ed.), 133–140. Amsterdam: Rodopi.
Kjellmer, Göran. 12–15 May, 2005. Collocations and semantic prosody. Paper presented at the American Association of Applied Corpus Linguistics, Ann Arbor MI.
Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104: 211–240.
Langacker, Ronald W. 2000. A dynamic usage-based model. In *Usage-based models of language,* M. Barlow & S. Kemmer (Eds), 1–63. Stanford CA: CSLI.
Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied Linguistics 18:* 141–165.
Leech, Geoffrey. 2000. Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50: 675–724.
Levelt, Willem J.M. 1989. *Speaking: From intention to articulation*. Cambridge MA: The MIT Press.
Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and technology. In honour of John Sinclair*, M. Baker, G. Francis & E. Tognini-Bonelli (Eds), Amsterdam: John Benjamins.
MacWhinney, Brian. (Ed.), 1999. *The emergence of language*. Hillsdale NJ: Lawrence Erlbaum Associates.

Meyer, David E. & Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90: 227–234.
Nattinger, James R. 1980. A lexical phrase grammar for ESL. *TESOL Quarterly* 14: 337–344.
Paivio, Allan. 1971. *Imagery and verbal processes*. New York NY: Holt, Rinehart and Winston.
Paivio, Allan. 1990. *Mental representations: A dual coding approach*. Oxford: OUP.
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), 191–225. London: Longman.
Pulvermüller, Friedemann. 1999. Words in the brain's language. *Behavioral and Brain Sciences 22: 253–336.*
Robinson, Peter & Nick C. Ellis (Eds), 2008. *A handbook of cognitive linguistics and SLA*. London: Routledge.
Schneider, Walte, Amy Eschman & Anthony Zuccolotto. 2002. *E-prime user's guide*. Pittsburgh PA: Psychology Software Tools.
Seidenberg, Mark S. 1997. Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275: 1599–1603.
Seidenberg, Mark S. & Maryellen C. MacDonald. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23: 569–588.
Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: OUP.
Sinclair, John. 2004. *Trust the text: Language, corpus and discourse*. London: Routledge.
Spivey, Michael. 2006. *The continuity of mind*. Oxford: OUP.
Tomasello, Michael. 2003. *Constructing a language*. Boston MA: Harvard University Press.
Warrington, Elizabeth K. 1975. The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* 27: 635–657.
Warrington, Elizabeth K. 1981. Neuropsychological studies of verbal semantic systems. *Philosophical Transactions of the Royal Society of London, B, B* 295: 411–423.

Frequency and the emergence of prefabs

Evidence from monitoring

Vsevolod Kapatsinski & Joshua Radicke
Indiana University

Abstract

Native English speakers were instructed to detect instances of /^p/ in spoken sentences by pressing a button as soon as they hear /^p/ regardless of whether it is inside another word. We observe that detection of the particle *up* is slower when the frequency of the verb + *up* collocation is low or extremely high than when it is medium. In addition, /^p/ is more difficult to detect in high-frequency words than medium-frequency or low-frequency words. Thus word frequency has a monotonic effect on detectability of word parts while the effect of phrase frequency is U-shaped. These results support the hypotheses that lexical units compete with their parts during speech perception and that words and ultra-high-frequency phrases are stored in the lexicon.

1. Introduction*

There is much evidence that language users are sensitive to co-occurrence statistics between words in both perception and production. In perception, Lieberman (1963) finds that predictable words are more intelligible. McDonald & Shillcock (2004) and Underwood et al. (2004), using eye-tracking, find that words that are probable given the preceding word or words are fixated for a shorter time than words that are not probable. Bod (2001) finds that subjects are faster in deciding that a three-word subject-object-verb sentence is grammatical when the sentence is frequent (*I like it*) than when it is not (*I keep it*). Reali & Christiansen (2007) present self-paced reading data that shows center-embedded relative clauses to be read faster when the embedded clause consists of a frequent pronoun-verb combination (*I liked*) than when it consists of an infrequent one (*I phoned*). Thus the frequency with which words co-occur (or some other co-occurrence statistic) must be stored in memory. The question we address is what effect frequent co-occurrence has on the memory representation of a pair of words.

One hypothesis, which we shall call **the distributed account**, is that co-occurrence simply increases the strength of an associative connection between the co-occurring words. Another hypothesis, **the localist account**, is that the co-occurring words fuse into a larger unit, the prefab, which has its own separate representation in memory (e.g., Bybee 2002; Wray 2002; Solan et al. 2005). This does not mean that the representations for the component words are lost as a result of the fusion. They may well be retained and even used during the production and perception of the frequent phrase. However, under the localist account, the prefab has its own node in the lexicon. That is, the prefab is a lexical unit, just like the words and morphemes that it contains. As Wray (2002: 265) puts it, a formulaic sequence is morpheme-equivalent.

Both theories can account for the finding that high-frequency phrases are processed more easily. In a high-frequency phrase, the end is somewhat predictable given the beginning and will therefore be easier to perceive. Sensitivity to predictability does not necessarily imply that the predictor and the predicted fuse into a unit. Rather, co-occurrence may simply make the co-occurring words able to prime each other.

However, in order to predict that high-frequency phrases are processed more easily than low-frequency phrases, the distributed account must predict

*Many thanks to Joan Bybee, Jill Morford, David Pisoni and Rena Torres-Cacoullos for helpful comments. Work supported by NIH training grant DC-00012 and NIH Research Grant DC-00111 to David Pisoni.

that the more predictable a word, the easier it is to process and detect (due to contextual priming). In particular, the final word of a frequent phrase should be perceived more easily than the final word of a less frequent phrase because the final word of a frequent phrase is predictable given the rest of the phrase and is primed by it.

This is not necessarily the case under a localist account in which prefabs are processed more easily (in part) because they are stored in the lexicon. The predictions of the localist account depend on how the processing of lexical units is hypothesized to interact with the processing of the units' parts. If one assumes that recognition of the whole helps with recognition of the parts (as, for instance, in the Interactive Activation Model of McClelland & Rumelhart 1981), then the localist account makes the same prediction as the distributed one (Healy 1994). If, on the other hand, recognition of the lexical unit interferes with processing of the unit's parts (Healy 1976), parts of high-frequency lexical units (i.e., prefabs) are predicted to be more difficult to detect than parts of low-frequency lexical units.

The idea of between-level competition during lexical access has been proposed independently by Healy (1976), Hay (2003) and Sosa & MacFarlane (2002). Corcoran (1966) and Healy (1976) observed more letter detection errors on the ultra-high-frequency word 'the' than on other words, e.g., the low-frequency word 'thy'. Furthermore, frequency has an effect even when grammatical class is controlled: letters are more difficult to detect in high-frequency nouns than in low-frequency nouns (Healy 1976; Minkoff and Raney 2000). Healy proposed the Unitization Hypothesis to account for the result:

> We can […] identify […] syllables, words, or even phrases, without having to complete letter identification. The identification of these higher-order units is facilitated by familiarity […] Once a larger unit is identified, the processing of its component letter units is terminated, even if the letters have not yet reached the point of identification. Instead, processing and attention are directed to the next location in the text. Because letter identification is not always completed for highly familiar words […] many letter-detection errors are made on these words. (Healy 1994: 333)

A limitation of the work using orthographic stimuli is that the results could be due to the fact that readers are less likely to fixate low-frequency words than high-frequency words during reading (Corcoran 1966; Inhoff & Rayner 1986). High-frequency words can be perceived parafoveally, where visual acuity is lower, which may impair the reader's ability to identify individual letters within words. Consistently with this interpretation, Hadley & Healy (1991) found that letter detection is no harder in *the* than in other words when subjects can view only five letters at once while reading text and thus are forced to fixate every word.

In the auditory modality, Sosa & MacFarlane (2002) found that detecting the word *of* in spoken sentences taken from the Switchboard Corpus was more difficult when *of* occurred in an ultra-high-frequency phrase such as *kind of* or *sort of* than when it occurred in a lower-frequency phrase, such as *couple of* or *think of*. No difference between medium-frequency and low-frequency collocations was found. Sosa & MacFarlane (2002) argue that extremely frequent phrases (prefabs) are stored in the lexicon and thus detecting *of* in them entails the extra step of morphological decomposition.

A limitation of Sosa & MacFarlane's study is that *of* undergoes much articulatory reduction in high-frequency collocations, such as *kind of* or *sort of*, often appearing without the consonant. This introduces a dilemma for investigating detectability of *of* in such phrases: if a reduced token of *of* is used, it is acoustically non-salient and difficult to perceive as well as being difficult to perceive as an instance of *of*. If a non-reduced token is used, then one is presenting the subject with an instantiation of *of* that is not typical for the context in which it appears. In either case, reaction times may be slowed down for reasons other than the collocation being stored as a single unit.

Thus, in the present study we asked subjects to monitor spoken sentences for a stimulus that does not show much articulatory reduction, the particle *up*. As Sosa and MacFarlane did with *of*, we examine the influence of the frequency of the prefab in which *up* occurs on how easy *up* is to detect. Based on Sosa and MacFarlane's results, we would expect *up* to be more difficult to detect when it occurs in a high-frequency verb+*up* combination like *sign up* than in a less frequent one like *pin up* or *run up*. Using *up* should allow us to test the idea that "it is frequency of use itself that determines the units of storage […]. The fact that the phrase is not (yet) reduced does not mean that it is not stored in memory as a unit" (Bybee 2001: 161). If high-frequency verb + *up* combinations are stored as lexical units, we would find evidence in support of the idea that abnormal phonological behavior is not a necessary precondition for storage.

Despite the fact that Sosa & MacFarlane did not find differences between low-frequency and medium-frequency phrases, there are reasons to suspect that *up* should be harder to detect in low-frequency phrases than in medium-frequency ones. Morton & Long (1976) and Dell & Newman (1980) found that phoneme detection was faster in words that were relatively predictable given the part of the sentence that preceded them relative to words that were not predictable, e.g., *book* vs. *bill* following *He sat reading a*; and *beer* vs. *brandy* following *He had a drink of* (from Morton & Long 1976). While at first glance this result appears to conflict with the results of Sosa & MacFarlane (2002), predictability of *beer* in *He had a drink of beer* is much lower than the predictability of *of* in *This was done kind of badly*. Conversely, *of* is still relatively predictable in the lowest-frequency

collocations used by Sosa & MacFarlane (2002), e.g., *sense of*, *piece of*, *each of*. Thus, existing evidence points to a U-shaped effect of phrase frequency on detectability of the phrase's parts: parts of a low-frequency phrase should be harder to detect than parts of a medium-frequency phrase which should be easier to detect than parts of an ultra-high-frequency phrase.

One type of model that predicts a U-shaped effect of phrase frequency on part detectability is one that assumes that a collocation is likely to be stored in the lexicon only if its frequency is above a certain threshold. This type of model has been advocated by Alegre & Gordon (1999) who did not find whole-word frequency effects for regularly inflected English words with a frequency below 6 per million while finding frequency effects throughout the frequency range for monomorphemic controls. If, like regularly inflected words in Alegre and Gordon's model, phrases are stored in the lexicon only if they are frequent enough and, other things being equal, predictability improves detectability, we should find facilitatory effects of predictability in phrases whose frequencies are insufficient for the phrase to become a stored prefab. One version of the theory is depicted in Figure 1.

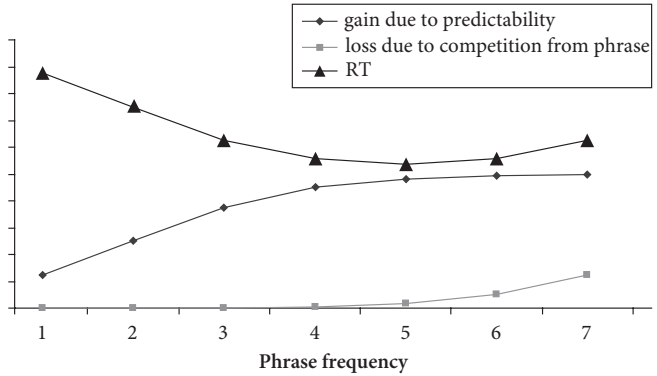


Figure 1. The theoretical relationship between phrase frequency and reaction time (RT) in detecting the second word in the phrase. Here $RT = A + loss - gain$ (predictability makes detection faster while competition from the prefab makes detection slower), where $Gain = \frac{1}{1+B^{2-PhraseFrequency}}$ while. $Loss = \frac{1}{1+B^{8-PhraseFrequency}}$.[1]

However, a U-shaped relationship between phrase frequency and word detectability is also expected in a model that assumes that the ease of detecting a word is a function of how easy it is to parse the word from the acoustic signal

(parseability) and how surprising, and therefore salient, the occurrence of the word is.[2] If the more predictable a word, the easier it is to parse from the signal, words in high-frequency phrases should be easier to detect than words in low-frequency phrases. However, at the same time, the occurrence of a word is not surprising if it is predictable and thus is less likely to attract attention, which could in turn lead to lower detectability. If, as phrase frequency increases, parseability rises faster than salience falls and parseability reaches ceiling (i.e., *up* is always parsed out) before salience reaches floor (i.e., the occurrence of *up* is not paid any attention at all), a U-shaped relationship between phrase frequency and word detectability is expected. Before parseability reaches the ceiling, detectability increases with increases in phrase frequency. After the ceiling is reached, salience is the only factor influencing detectability, hence further increases in phrase frequency should decrease word detectability.

In order to distinguish between the two theories, we need to look at what happens when parseability is not at ceiling and when wholes at the low end of the frequency continuum are also likely to be stored. This can be accomplished by looking at stimuli in which the to-be-detected stimulus, /^p/, is not a word but instead occurs inside a word, e.g., *puppy*. In these cases, *up* is less likely to be parsed from the signal and parseability is not at ceiling (accuracy in *up* detection is not perfect). Hence, inhibitory effects of ultra-high-frequency should not be found for word-internal /^p/s if they are due to a parseability/salience tradeoff.

On the other hand, if the decrease in parseability of the parts is due to increased competition from the whole, /^p/ should be harder to detect in high-frequency words than in low-frequency words. Furthermore, since all words we examine are likely to be stored in the lexicon, there should be a negative correlation between /^p/ detectability and word frequency throughout the frequency range.

1. B and A are constants. The crucial feature is that the power to which B is raised is larger in the Loss formula than in the Gain formula. A processing interpretation of this mathematical formulation of the theory is that the word and the prefab are nodes with a sigmoid activation function. During recognition, the prefab and its parts compete for a limited amount of activation where the amount of activation received by a node is proportional to its resting activation level. The constant A represents the minimum time required to make a detection response.

2. This is Corcoran's (1966) idea that predictable words are skipped over/not attended to generalized to auditory perception.

2. Methods

2.1 Materials

The verb + *up* collocations were chosen for inclusion in the experiment based on having non-zero frequency in the British National Corpus (determined through the online interface at http://view.byu.edu/). The British National Corpus was chosen because of its size and the availability of part-of-speech tagging. To find all verb + *up* constructions, we searched for the following pattern: [v*] up.[avp]. We obtained the frequencies of the verb + *up* collocations from the corpus.

The final sample of collocations used in the study was derived by keeping the 10 collocations closest to each end of the frequency continuum and randomly sampling the remaining collocations. In addition, we took all verbs that occurred with the particle *out* in the corpus and included a sample of such verbs that did not occur with *up* in the corpus but did occur with it on Google (the least frequent of these was *eke up*, as in *Tokyo's Nikkei slipped 0.9% and the FTSE 100 in London eked up 0.1%*.) paired with *up* to create the ultra-low-frequency end of the frequency distribution where *up* is not very predictable.

Most of the verb-particle phrases were presented using the past tense form of the verb. For regular verbs, this ensured that *up* was preceded by /d/ or /t/ (sometimes a flap). This was done to ensure that the location of the vowel onset in *up* can be reliably measured and to minimize the influence of phonological context on detectability of *up*.

The first author created 240 experimental sentences containing the particle *up* and 240 control sentences that were identical to the experimental sentences except for containing a different particle. The sentences were presented to the second author, a native English speaker, in a randomized order. The second author read the sentences aloud, having a fixed amount of time (5 seconds) to produce each sentence.

Thirty-five of the control sentences contained the particle *out*. Since experimental and control sentences were syntactically identical, prosody was not a cue to whether *up* occurs in the sentence. In most sentences, *up* was located immediately after the verb. However, to ensure that the subjects process the entire sentence, there were control sentences in which *up* either followed the direct object (*He brought it up*) or was sentence-initial (*Up he goes*). A verb occurring in these control sentences also occurred in an experimental sentence. The control sentences containing *up* were paired with control sentences of the same syntactic structure that contained a different particle so that the number of sentences containing *up* was equal to the number of sentences not containing *up*. The control sentences in which *up* is not immediately after the verb are not included in the analyses presented in this paper because the frequency of verb+*up* combinations was

determined only for the most frequent location of *up*, which is immediately after the verb. The subject of the sentence was almost always a pronoun to ensure lack of co-occurrence-based priming between the subject and the particle. Twenty sentences containing noun-phrase subjects occurred in both the experimental and the control set to increase variability in particle location. Previous research has suggested that the greater the variability in location of the to-be-detected unit, the greater the likelihood of obtaining context effects (Lively & Pisoni 1990)

In addition to stimuli in which *up* is a particle, we included a set of sentences in which /^p/ was inside another word. These sentences increase variability in target location and allow us to examine how word frequency influences detectability of parts of the word. We can then compare the influence of word frequency to the influence of phrase frequency. The words used were found in the MRC Psycholinguistic Database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm, Coltheart 1981). For the experimental sample, we excluded compounds (e.g., *buttercup*), verb-particle constructions, words in which /^p/ was followed by a stop (e.g., *interrupt*), and Internet terms, whose frequency would be elevated in Google counts relative to overall use (*pop-up*, *lookup*, *setup*). We did not exclude nouns and adjectives derived from verb-particle constructions (e.g., *holdup*). If a noun could be used in the plural, we created two sentences, one containing the noun in the plural and one containing it in the singular.

It was ensured that /^p/ was equally likely to occur word-finally (e.g., *holdup*, *cup*), word-medially (e.g., *puppy*, *hiccups*) and word-initially (e.g., *upholstery*, *upper*). Morphological and syllabic constituency of /^p/ was manipulated. For instance, /^p/ is a syllabic constituent (the rime) but not a morphological constituent in *cup* while it is a morphological constituent that crosses a syllable boundary in *upper*. There were 96 /^p/-containing words used in the experiment. Each sentence with an /^p/-containing word was paired with a control sentence in which the /^p/-containing word was replaced by a word containing /aʊt/. The /aʊt/-containing words were also found using the MRC Psycholinguistic Database using the same exclusion criteria as for /^p/-containing words.

2.2 Subjects and procedure

Twenty adult native English speakers were recruited from among introductory psychology students. They participated to fulfill a course requirement. The subjects were asked to press the 'present' button as soon as they hear *up*, regardless of whether it is a separate word or is inside another word. If the sentence did not contain *up*, they needed to press the 'absent' button to go on to the next sentence. They were encouraged to respond as soon as they hear *up* without waiting until the end of the sentence. The experiment lasted approximately 25 minutes.

2.3 Measurement of frequency and duration

For the purposes of deriving frequency-detectability correlations, we obtained phrase frequency estimates from the spoken portion of the British National Corpus (BNC) and Google. While a U-shaped phrase frequency- word detectability relationship was observed with both counts, the Google-based results exhibited both a larger facilitatory effect on the low-frequency end of the continuum and a larger inhibitory effect at the high-frequency end. Furthermore, the spoken portion of the BNC did not allow us to distinguish between many frequency classes at the low-frequency end of the continuum. Thus only Google results are reported in this paper.

The use of web-based frequency estimates of phrase frequency is supported by the results of Keller & Lapata (2003) who found that plausibility judgments for bigrams that are found only on the Web (and not in the BNC) are reliably predicted by Google frequencies, indicating that Google counts are capturing psychologically relevant variation on the low end of the phrase frequency continuum that the BNC counts are not. Furthermore, even for bigrams found both in the BNC and on Google, correlations with plausibility judgments were higher for web-based frequency counts than for corpus-based ones.

Both base and surface frequency estimates were derived. The surface frequency estimate is the frequency of the verb + *up* combination where the verb is in the particular inflected form used in the experiment. The base frequency estimate is the summed frequency of verb + *up* summed across all forms of the verb. The results did not differ depending on whether base or surface frequency estimates were used.

In analyzing the effect of phrase frequency, the frequency continuum was split into seven bins based on natural discontinuities in our sample of frequencies, as shown in Figure 2.

To investigate the effect of phonological reduction on detectability, we measured the durations of each occurrence of *up* in the materials. We also measured the distance between *up* and the beginning of the sentence. All measurements were done in Praat. The release of the stop closure was taken as the end of the particle. Following stops and fricatives, the beginning of the particle was determined by the beginning of the vowel formants on the spectrogram (since the preceding verb was almost always in the past tense, this was the usual case). When the vowel onset was not readily apparent on the spectrogram, we listened for cues to the identity of the vowel in the preceding speech signal. We took the onset of the vowel to be the latest point at which we could not yet detect cues to the identity of the upcoming vowel. In order to control for possible effects of phonological reduction and measurement error, we measured reaction time both from the onset and the offset of the particle.

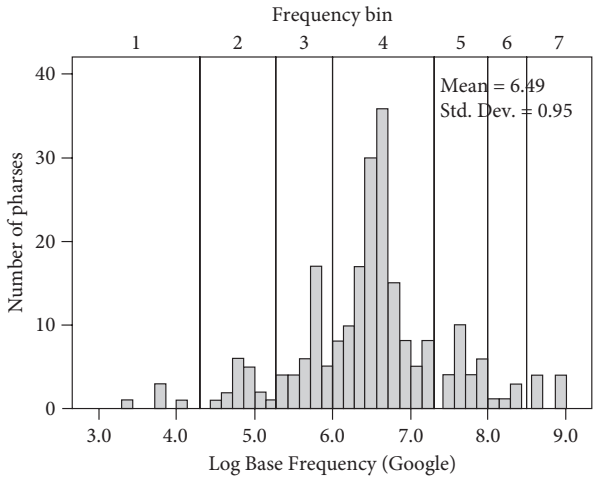


Figure 2. The frequency bins were derived based on discontinuities in the sample of frequencies.

3. Results

3.1 /^p/ as a particle

Unlike in Sosa & MacFarlane (2002), accuracy in particle detection in the present study was quite high. Sosa & MacFarlane report that accuracy of *of* detection was at 47% in the lowest-frequency phrases, 60% in medium-low-frequency phrases, 38% in medium-high-frequency phrases, and 37% in the ultra-high-frequency phrases. Results from the present experiment are shown in Table 1. Accuracy in the lowest-frequency group is significantly lower than in any other group (with all other groups combined $p < .0005$; according to one-way ANOVA). Frequency bins 5 and 6 exhibit higher accuracy than either bin 7 ($p = .038$), or bins 2, 3, and 4 ($p = .005$). These results indicate that *up* is easier to detect when it is somewhat predictable than when it is unexpected (Morton & Long 1976; Dell & Newman 1980). The data suggest a U-shaped relationship with accuracy steadily increasing with phrase frequency but then dropping for the highest-frequency bin.

Table 1. Error rate in up detection depending on the frequency of the verb + up collocation

frequency bin	1 lowest	2	3	4	5	6	7 highest
error rate	20%	5%	6%	5%	3%	2%	6%

Figure 3 presents reaction time (**RT**) data (correct trials only). As predicted by the hypothesis of between-level competition between prefabs and their component words, detection of *up* is more difficult in ultra-high-frequency verb + *up* collocations than in medium-frequency collocations. The difference in reaction time between frequency bin 7 (the highest-frequency bin containing the collocations *get up*, *sign up*, *go up*, and *set up*) and bin 6 (containing slightly less frequent collocations, including *keep up*, *line up*, *stand up*, *catch up*) is statistically significant according to a one-way ANOVA (for reaction time relative to particle onset, $p = .005$, for reaction time relative to particle offset, $p = .002$). Interaction with subject identity is not significant ($p > .1$). The significance of this effect is further confirmed by the fact that a quadratic function, which is U-shaped, provides a much better fit to the data than a monotonic, logarithmic one (the quadratic function explains 96% of the variance in reaction time as a function of phrase frequency while the logarithmic function explains 57% of the variance in reaction time measured relative to the onset and 46% of the variance in reaction time relative to the offset). The effect is observed regardless of whether we estimate phrase frequency via base frequency or surface frequency (for surface-frequency estimates, the difference between groups 7 and 6 is significant at $p<.05$, while the difference between groups 7 and 5 is significant at $p = .002$, interactions with subject identity are not significant, $p > .2$).

The difference in fit almost disappears if frequency bin 7 is removed (the fit of the logarithmic function increases to 94–95% of the variance) indicating that

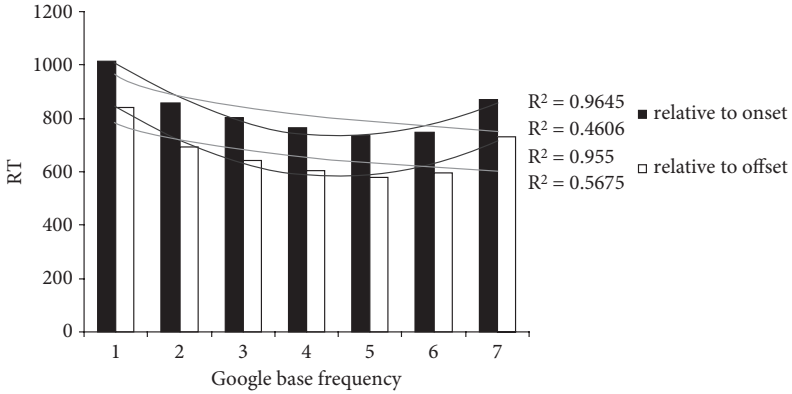


Figure 3. The U-shaped effect of the frequency of verb + up collocations on the speed with which ***up*** is detected. For both RT measured from the beginning of the word and RT measured from the end to the word, the top R^2 value indicates the amount of variance accounted for by the U-shaped function while the bottom R^2 value indicates the amount of variance accounted for by the monotonic function.

throughout most of the frequency range, increased predictability helps to detect the particle. Just like in Sosa & MacFarlane (2002) and consistent with the accuracy results above, effects of phrase-word competition are only observed with extremely high-frequency phrases. Throughout most of the frequency continuum, *up* detection is easier in higher-frequency phrases than in lower-frequency ones, supporting the hypothesis that, other things being equal, predictability of the to-be-detected unit speeds up detection (Morton & Long 1976; Dell & Newman 1980).

In order to examine how consistent our results are with the results of Sosa & MacFarlane (2002), we examined where the collocations used in the previous study fit onto the frequency continuum derived from Google. We obtained a mean log frequency of 8.15 for their lowest-frequency group, 8.36 for the medium-low-frequency group, 8.77 for the medium-high-frequency group and 8.92 for the ultra-high-frequency group. Thus, their lowest-frequency bin is similar in frequency to our bin 6 (mean log frequency = 8.22) while our group 7 is similar to their medium-high-frequency group (mean log frequency = 8.72). Thus, we find the inhibitory frequency effect at a similar (slightly lower) frequency level than Sosa & MacFarlane. The absence of facilitatory predictability effects in Sosa and MacFarlane's data is consistent with our findings: such effects are found much lower on the frequency continuum (between bin 1 with mean frequency of 3.74 and bin 5 with mean frequency of 7.72) than the range of frequencies used by Sosa & MacFarlane.

Importantly, the duration of the particle does not depend on phrase frequency. As can be seen in Figure 3, the difference between reaction time relative to particle onset and reaction time relative to particle offset is constant throughout the frequency range. Thus, the slow-down in detection observed in ultra-high-frequency phrases is not due to the presence of phonological reduction in those phrases. Thus, the findings of the present study support the hypothesis that phonological reduction is not a precondition for storage (Bybee 2001).

3.2 Word-internal /^p/

An alternative interpretation of the results in the previous section is a parseability-salience tradeoff: at some point on the phrase frequency continuum, *up* becomes so predictable that it is always parsed out of the signal. Above that point, further increases in phrase frequency can only decrease how surprising the occurrence of *up* is without increasing the likelihood of *up* being parsed out. To test this hypothesis, we turn to data from trials in which /^p/ occurs inside another word. In such cases, parseability of /^p/ should be decreased, thus /^p/ may be easier to detect in high-frequency words than in low-frequency words. On the other hand, since words are stored in the lexicon, the hypothesis of between-level competition

predicts that /^p/ should be harder to detect in high-frequency words because such words are stronger competitors. A U-shaped function is not predicted because even the lowest-frequency words are expected to be stored in the lexicon.

Since word-internal occurrences of /^p/ are not all equal in terms of location within the word, length of the bearing word, morphological and syllabic constituency, stress, and, as it turns out, duration, we tested for effects of each of these variables. While stress and within-word location did not have a significant main effect, morphological and syllabic constituency, word length, and duration did.

Table 2. shows that /^p/ is easier to detect when it is a morpheme than when it is not ($p < .0005$ for both accuracy and reaction time). This result is consistent with Zwitserlood et al.'s (1993) findings for syllable monitoring in Dutch.

Table 2. /^p/ is easier to detect when it is a morpheme than when it is not[3]

	Morpheme	Not morpheme
Accuracy	90%	72%
Reaction time	813	1023

As shown in Table 3, accuracy of /^p/ detection is also affected by the length of the word in which /^p/ occurs: /^p/ is more likely to be missed in longer words than in shorter ones ($p = .002$ in a multinomial logistic regression that also included morphological constituency, syllabic constituency, and presence/absence of stress) especially if /^p/ is not a morpheme (the interaction is significant at $p = .026$). Table 3 shows that this is not a side effect of differences in duration of /^p/ within long and short words: while in general, longer instances of /^p/ are easier to detect (Table 6), instances of /^p/ that occur in longer words do not tend to be shorter than those occurring in short words (in fact, instances of /^p/ tend to be somewhat longer in longer words).

Table 3. The effect of word length on accuracy of /^p/-detection (number of segments by percent correct)

Length (segments)		3	4	5	6	7	8	10
% correct	Morpheme	N/A	95%	92%	90%	87%	86%	N/A
	Not morpheme	88%	76%	73%	58%	55%	N/A	55%
duration of /^p/ (ms)	Morpheme	N/A	93	94	99	102	116	N/A
	Not morpheme	74	64	84	134	112	N/A	47

3. Reaction time for word-internal occurrences of /^p/ is relative to the onset of /^p/.

The effect of word length is consistent with the hypothesis of between-level competition. There is a greater chance that not all parts of a word will be fully perceived prior to word identification in a long word than in a short word. Thus, processing of a part is more likely to be interrupted prior to completion in a long word than in a short word. If this hypothesis is correct, then, given that words are processed mostly left-to-right, the effect of word length should be most apparent in the word-final position, less apparent in the word-medial position and least apparent in the word-initial position. This is indeed the case in the data: the effect of word length is highly significant in the word-final position according to a one-way ANOVA ($p < .0005$ for non-morphemic and $p = .008$ for morphemic /^p/'s), marginally significant in the word-medial position ($p = .087$ for non-morphemic and $p = .063$ for morphemic /^p/'s), and not significant in the word-initial position ($p = .172$ for non-morphemic and $p = .186$ for morphemic /^p/'s).

Table 4 shows that detection of /^p/ is slower when /^p/ straddles a syllable boundary than when it does not ($p < .0005$). There was no difference between cases in which /^p/ is a syllable and when it is the rime (whether or not the rime was followed by an appendix). Syllabic constituency does not have a significant effect on accuracy, although the numerical trend is in the same direction as the effect on reaction times (87% correct when /^p/ is a syllabic constituent vs. 85% when it straddles a syllable boundary).

Table 4. The effects of morphological and syllabic constituency on the speed of /^p/ detection (ms)

	Morpheme	Not a morpheme
Syllabic constituent	796	960
Not a syllabic constituent	964	1187

The effect of syllabic constituency on sequence monitoring has been previously obtained by Mehler et al. (1981) for French, Bradley et al. (1993) for Spanish, and Zwitserlood et al. (1993) for Dutch. It has not previously been found in English (Cutler et al. 1986; Bradley et al. 1993). A possible reason for why previous studies have not found a syllabic constituency effect is that both Cutler et al. (1986) and Bradley et al. (1993) had subjects monitor for sonorant-final targets[4] whereas we used a stop-final target. A post-vocalic sonorant in English is more closely associated with the preceding vowel than an intervocalic stop is (Treiman & Danis 1988; Derwing 1992). Thus, previous syllable monitoring studies in English may not have included (many)

4. Cutler et al. (1986) used /l/, Bradley et al. (1993) used mostly /l/ and nasals except for two stimuli containing /s/.

targets that crossed a syllable boundary. This hypothesis is supported by the results of Ferrand et al. (1997) who failed to observe an effect of prime-target syllable structure consistency in masked priming in English when using Bradley et al.'s (1993) stimuli but were able to obtain it when stimuli with clear syllable boundaries were used.

The findings in Tables 2–4 indicate that /^p/ is more detectable when it is a constituent (whether morphological or phonological) than when it is not. These findings support a view of constituency as unithood: constituents are more likely to be parsed out of the signal than phoneme strings that straddle a constituent boundary. Especially in longer words, not all parts of the word are parsed out of the signal. Being a constituent makes a phoneme string more likely to be detected.

There is no interaction between morphological and syllabic constituency for either accuracy or reaction time ($p > .3$), indicating that being a syllabic constituent increases detectability even when /^p/ is a morphological constituent. Similarly, being a morpheme increases detectability of units that are syllables or rimes. This suggests that a morphological or syllabic constituent is not always parsed out of the signal. Rather, the fewer the constituent boundaries that lie within a phoneme string, the more likely the string is to be parsed out.

However, before we conclude that constituency affects detectability, we need to address the fact that constituency of the particle correlates with particle duration in the stimuli, as shown in Table 5. Main effects of morphological and syllabic constituency are significant ($p < .0005$ in an ANOVA that included morphological constituency, syllabic constituency and word length as fixed factors and subject as random factor). There is no significant interaction.

Table 5. The effect of constituency on duration of /^p/ (ms)

	Morpheme	Not a morpheme
Syllabic constituent	100	86
Not a syllabic constituent	84	67

There is a significant correlation between /^p/ duration and how easy it is to detect. Shorter, more reduced, instances of /^p/ are detected more slowly (Pearson $r = -.27$, $p < .0005$).[5] Therefore, we conducted a linear regression analysis with logarithmically scaled reaction time as a dependent variable and syllabic constituency (1 vs. 0), morphological constituency (1 vs. 0), presence of stress on /^p/, /^p/ duration, word length (in segments), distance from sentence onset to /^p/ onset, log word frequency, and location of the stimulus in the list of sentences as independent variables. Both of the constituency variables were significant ($t = -4.123$, $p = .001$ for syllabic constituency, $t = -3.227$, $p < .0005$ for morphological

5. We used log10(reaction time) for correlation analyses.

constituency) as was duration of /^p/ ($t = -4.206$, $p < .0005$). These results suggest that constituency has an effect on detectability above and beyond duration.

In this analysis, the effect of word frequency only approached significance ($p = .089$, $t = 1.702$). The direction of the trend was as predicted by the hypothesis of between-level competition: /^p/ was more difficult to detect in high-frequency words than in low-frequency words. However, we reasoned that the word frequency effect may not manifest itself when /^p/ occurs in the word-initial position but only when /^p/ occurs word-medially or word-finally. For instance, Lively & Pisoni (1990) observe a much stronger word frequency effect in phoneme categorization when the phoneme was in the final position than when it was in the initial position of a CVC word. In addition, we have observed earlier that the effect of word length on detectability of the word's parts is stronger for non-initial parts.

Thus, we broke the data down by where in the word /^p/ was located. Table 6 shows correlations between /^p/ duration, log frequency and logarithmically scaled reaction time depending on where in the word /^p/ is located. All correlations are significant ($p < .001$) except the one between word frequency and reaction time in the word-initial position, indicating that while word frequency does not appear to affect detection of word-initial targets, this is not simply because word-initial data is messier. The correlations between word frequency and speed of /^p/ detection are in the direction predicted by the between-level competition hypothesis: the higher the frequency of the word, the harder /^p/ is to detect when it occurs inside it.

Table 6. Correlations (r) between independent variables and reaction time to /^p/ depending on the location of /^p/ within the word

	Initial	Medial	Final
Word frequency	.052	.285	.221
/^p/ duration	–.264	–.231	–.282

When word-initial instances of /^p/ are excluded from the regression analysis, word frequency is a significant predictor of reaction time ($t = 2.999$, $p = .003$). Figure 4 shows that when a variety of functions is fit to the data, all of them display a monotonic relationship between word frequency and reaction time. Thus as word frequency increases, time taken to detect /^p/ inside the word rises throughout the frequency range. Unlike the effect of phrase frequency, the effect of word frequency is not U-shaped, as expected if (1) all words we presented to subjects are stored in the lexicon, (2) lexical units compete with their parts during recognition, and (3) high-frequency lexical units are stronger competitors.

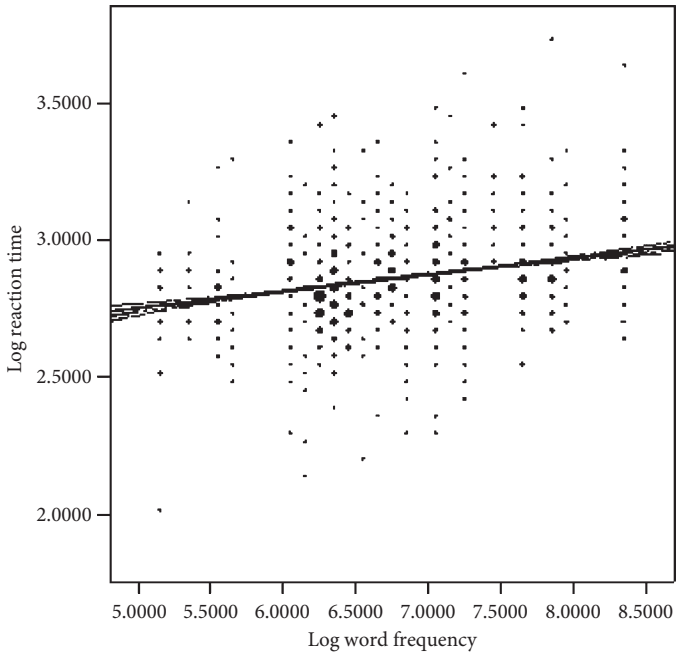


Figure 4. The monotonic relationship between word frequency and detectability of /^p/ within the word.[6]

3.3 Summary of the results

When *up* is a particle:

1. The higher the frequency of the verb-particle collocation, the easier the particle is to detect, except for the highest-frequency collocations.
2. Detection of the particle is harder in the highest-frequency verb-particle collocations than in less frequent collocations.

When /^p/ is inside another word and is not word-initial:

3. The higher the frequency of the word, the harder it is to detect /^p/ inside it.
4. The longer the word, the harder it is to detect /^p/ inside it.

6. Circle size indicates number of data points. The trendlines shown are linear, quadratic, cubic and sigmoid.

Regardless of whether /^p/ is word-initial:

5. /^p/ is harder to detect when it crosses a morphological or syllabic constituent boundary than when it is a morphological or syllabic constituent.
6. Short instances of /^p/ are harder to detect than longer instances.

4. Discussion

4.1 Theoretical interpretation

The phoneme sequence /^p/ is more difficult to detect inside a high-frequency word than inside a low-frequency word. Thus, parts of frequent lexical units are less accessible to detection than parts of rare lexical units. Given this finding, we would predict that, if prefabs are lexical units, parts of frequent prefabs should be harder to detect than parts of rare prefabs. Finding an inverse relationship between frequency of a whole and detectability of its parts should indicate that at least the high-frequency wholes are stored in the lexicon. Such an inverse relationship is found for verb-particle phrases containing *up* but only at the very top of the phrase frequency continuum. These results are consistent with Sosa & MacFarlane's (2002) findings on word+*of* collocations. They indicate that the highest-frequency phrases are stored in memory as lexical units but they also **suggest** that a phrase needs to be extremely frequent to be stored in the lexicon.[7]

Why are parts of high-frequency lexical units harder to detect than parts of less frequent lexical units? There must be some mechanism that would make activating the prefab interfere with bottom-up activation of the component words and activating a word interfere with bottom-up activation of the component morphemes, syllables, and bigrams. In other words, the results can only be explained if linguistic units in a part-whole relationship compete for activation during the perception process. This hypothesis is also supported by our finding that /^p/ is more likely to be missed in a long word, where recognition of /^p/ is less likely to be necessary for lexical access.

7. However, as Figure 1 shows, it is also possible that the activation level of the phrase begins to rise slowly as phrase frequency increases, and that until a certain point these frequency-dependent increases in the amount of competition the phrase generates are not enough to offset increases in word predictability that are also caused by increases in phrase frequency. If that is the case, a more prudent conclusion is that the phrase representation does not participate in the lexical access process to a significant degree unless the phrase is extremely frequent.

This idea can be implemented in several non-mutually-exclusive ways. Some possibilities include (1) competition for a limited supply of activation coming from either the acoustic signal or previously perceived context, (2) top-down inhibition, where wholes inhibit their parts when activated beyond a particular threshold (Libben 2005: 276), or (3) removal of the activation source at the completion of lexical access by ceasing to process the acoustic signal that has been parsed into lexical units (Healy 1994).

Finally, we observe that /^p/ is easier to detect when it is a constituent than when it is not a constituent. This finding suggests that the acoustic signal is parsed into morphemes and syllables during speech perception making /^p/ easier to detect when it matches one of the units automatically extracted from the signal and more difficult to detect when the component segments of /^p/ need to be matched to segments that occur in different, though adjacent, units.

4.2 The facilitatory effect of word frequency on phoneme monitoring in word lists

In the present study, we observed that sequence detection is easier in low-frequency words than in high-frequency words. This is consistent with letter-detection results observed by Healy (1976) and Minkoff & Raney (2000). However, a word frequency effect in the opposite direction is often observed in phoneme monitoring (Rubin et al. 1976; Cutler et al. 1987; Eimas et al. 1990; Lively & Pisoni 1990) and letter monitoring (Howes & Solomon 1951; Johnston 1978) where phonemes and letters in high-frequency words are easier to detect than those in low-frequency words.

There is a systematic difference between experiments that find a word-frequency advantage in letter or phoneme detection and those that find a disadvantage: the word-frequency advantage is found with single-word presentation while multi-word presentation yields a word-frequency disadvantage (Healy et al. 1987; Hadley & Healy 1991).[8]

Healy et al. (1987) explain the difference between single-word and multi-word presentation using the Unitization Hypothesis. According to the hypothesis, readers move on to the next word in text as soon as they have identified the current word, terminating processing of smaller units within the current word. When only a single word is visible, there is no subsequent word, hence the subjects will continue processing the word they have already identified, at which point determining the identity of individual letters will be facilitated by having identified the word

8. Eimas et al. (1990) presented target words in a sentence context but the context was constant (the next word is…) and the target word was always the last word in the sentence.

because the reader will be able to use his/her knowledge of what the word is to infer whether the target letter has been presented.

This explanation predicts that the word-frequency disadvantage should not be observed when the target word is in the sentence-final position. Our data are consistent with this prediction: there is no significant correlation between log word frequency and log reaction time for words in the sentence-final position even if only words in which /^p/ is not word-initial are included ($r = .047$, $p = .569$). However, this subset of words is small (12 words), so the reliability of this result is questionable.

5. Conclusion

Listeners find it more difficult to detect /^p/ in a high-frequency lexical unit than in a low-frequency one or, more concisely, **the stronger the whole the weaker the parts** (Bybee & Brewer 1980; Hay 2003; Healy 1976; Sosa & MacFarlane 2002). While all words are lexical units, leading to a monotonic relationship between word frequency and difficulty of /^p/ detection, our results suggest that only high-frequency phrases are stored in the lexicon. Since, other things being equal, predictable units are easier to detect, there is a U-shaped relationship between the frequency of the verb-particle collocation and detectability of the particle. For collocations that are not stored in the lexicon as units, the more probable the particle, the easier it is to detect due to a strong association between the particle and the co-occurring verb. For phrases that are stored in the lexicon, the more frequent the phrase, the more it interferes with the detection of the particle. Finally, /^p/ is easier to detect when it matches a morphological or syllabic constituent than when the segments of /^p/ are separated by a morpheme or syllable boundary, providing evidence for the hypothesis that syllables and morphemes are extracted from the acoustic signal and take part in the part-whole competition operating during lexical access.

References

Alegre, Maria & Peter. Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40: 41–61.

Bod, Rens. 2001. Sentence memory: The storage vs. computation of frequent sentences. Paper presented at the CUNY Sentence Processing Conference. Philadelphia PA.

Bradley, Dianne C., Rosa M. Sánchez-Casas & Juan E. García-Albea. 1993. The status of the syllable in the perception of Spanish and English. *Language and Cognitive Processes* 8: 197–233.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In *The evolution of language out of pre-language,* T. Givón & B.F. Malle (Eds), 109–32. Amsterdam: John Benjamins.

Bybee, Joan. 2001. *Phonology and language use*. Camdridge: CUP.

Bybee, Joan & Mary A. Brewer. 1980. Explanation in morphophonemics: Changes in Provençal and Spanish preterite forms. *Lingua* 52: 201–42.

Coltheart, Max. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology* 33A: 497–505

Corcoran, Derek. W.J. 1966. An acoustic factor in letter cancellation. *Nature* 210: 658.

Cutler, Anne, Jacques Mehler, Dennis G. Norris & Juan Segui. 1987. Phoneme identification and the lexicon. *Cognitive Psychology* 19: 141–77.

Cutler, Anne, Jacques Mehler, Dennis G. Norris & Juan Segui. 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* 25: 385–400.

Dell, Gary S. & Jean E. Newman. 1980. Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior* 19: 607–23.

Derwing, Bruce L. 1992. A 'pause-break' task for eliciting syllable boundary judgments from literate and illiterate speakers: Preliminary results for five diverse languages. *Language and Speech* 35: 219–35.

Eimas, Peter D., Susan B. Marcovitz-Hornstein & Paula Payton. 1990. Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language* 29: 160–80.

Ferrand, Ludovic, Juan Segui & Glyn W. Humphreys. 1997. The syllable's role in word naming. *Memory and Cognition* 25: 458–70.

Hay, Jennifer. 2003. *Causes and consequences of word structure*. London: Routledge.

Hadley, Jeffrey A. & Alice F. Healy. 1991. When are reading units larger than the letter? Refinement of the unitization reading model. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17: 1062–73.

Healy, Alice F. 1994. Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin and Review* 1: 333–44.

Healy, Alice F. 1976. Detection errors on the word *the*: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance* 2: 235–42.

Healy, Alice F., William L. Oliver & Timothy P. MacNamara. 1987. Detecting letters in continuous text: Effects of display size. *Journal of Experimental Psychology: Human Perception and Performance* 9: 413–26.

Howes, Davis H. & Richard L. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology* 41: 401–10.

Inhoff, Albrecht W. & Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics* 40: 431–9.

Johnston, James C. 1978. A test of the sophisticated guessing theory of word perception. *Cognitive Psychology* 10: 123–53.

Keller, Frank & Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29: 459–84.

Libben, Gary. 2005. Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics* 50: 267–83.

Lieberman, Philip. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6, 172–87.*

Lively, Scott E. & David B. Pisoni. 1990. Some lexical effects in phoneme categorization: A first report. *Research on Speech Perception Progress Report* 16: 327–59. Bloomington IN: Indiana University Speech Research Lab.

McClelland, James L. & David E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review* 88: 375–407.

McDonald, Scott A. & Richard C. Shillcock. 2004. Eye-movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science* 14: 648–52.

Mehler, Jacques, Jean-Yves Dommergues, Uli Frauenfelder & Juan Segui. 1981. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior* 20: 298–305.

Minkoff, Scott R.B. & Gary E. Raney. 2000. Letter-detection errors in the word *the*: Word frequency versus syntactic structure. *Scientific Studies of Reading* 4: 55–76.

Morton, John & John Long. 1976. Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior* 15: 43–51.

Reali, Florencia & Morton H. Christiansen. 2007. Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology* 60: 161–70.

Rubin, Philip, Michael T. Turvey & Peter Van Gelder. 1976. Initial phonemes are detected faster in words than in non-words. *Perception and Psychophysics* 19: 394–8.

Solan, Zach, David Horn, Eyton Ruppin & Shimon Edelman. 2005. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences* 102: 11629–34.

Sosa, Anna V. & James MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language* 83: 227–36.

Treiman, Rebecca & Catalina Danis. 1988. Syllabification of intervocalic consonants. *Journal of Memory and Language* 27: 87–104.

Underwood, Geoffrey, Norbert Schmitt & Adam Galpin. 2004. The eyes have it: An eye-movement study into the processing of formulaic sequences. In *Formulaic sequences.* Acquisition, processing and use [Language Learning & Language Teaching 9], N. Schmitt (Ed.), 153–72. Amsterdam: John Benjamins.

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

Zwitserlood, Pienie, Herbert Schriefers, Aditi Lahiri & Wilma van Donselaar. 1993. The role of syllables in the perception of spoken Dutch. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19: 1–12.

PART III

Functional explanations

Formulaic argumentation in scientific discourse

Heidrun Dorgeloh & Anja Wanner
Universität Düsseldorf & University of Wisconsin-Madison

Abstract

In this paper we analyze linguistic strategies of constructing an argument in scientific texts, in particular a formulaic construction that we call the *paper* construction, as in *This paper argues*…. Based on a corpus of 160 abstracts from different academic disciplines, we find that formulae of this kind are not a marginal phenomenon, but are preferred today in certain contexts over both the passive and the agentive active. A corpus study based on historical data from the ARCHER corpus indicates that the construction is a rather recent phenomenon. We offer a hypothesis about the historical predecessor of this construction and relate the rise of these formulae to a rhetorical shift from presenting scientific evidence as discoveries to a more constructional approach.

1. Introduction

In this paper we examine an instance of formulaic language that we call the *paper* construction. This construction combines active morphology with an inanimate subject in that the person who is bringing forward an argument is replaced by a textual category. This results in sequences such as *This paper argues…* or *This article analyzes…*, which in principle violate the selectional restrictions of the verb, which would normally require an agent as its subject.[1] Semantic and/or syntactic irregularities are one criterion for considering the "recurrence" of the "co-occurrence of words" (Gledhill 2000: 8) as formulaic (Cowie 1992; Wray & Perkins 2000). In addition, what licenses them in this case is the context of the genre in which they occur: We will see that the *paper* construction has been brought about by genre-specific changes in what counts as a good or a relevant argument in science and that it has been contingent on the development of specific genres, in particular research articles and their accompanying abstracts. Following Gledhill (2000: 206), the construction can therefore also count as an instance of "generic collocation".

In section 2 of this paper we will discuss characteristics of scientific English in general and of the abstract as a genre in particular. Section 3 will be concerned with the role of constructing an argument in abstracts and with typical collocations for realizing such argumentation. Our observations are based on a synchronic corpus of 160 abstracts from scholarly articles published in journals from different academic fields,[2] but we will also discuss evidence from a diachronic study of written scientific English, based on the ARCHER corpus.[3] Data from both studies are presented in section 4; they indicate that collocations involving a *paper*-like subject are not a peripheral phenomenon, especially not in certain academic disciplines: While they are less prominent in the natural sciences, they have become the most common strategy for presenting an argument in the humanities. In section 5, we discuss these results in more detail. We show that, in view of the pragmatics of

1. Formulae of this kind are also in line with the observation that scientific English scores high on abstract, inanimate subjects and low on agentivity (e.g., Biber 1988). A related case in point is discussed by Kerz & Haas (this volume), who show that typical research predicates are part of more complex and partly nominal constructions in academic discourse.

2. Our data are from the following journals: *American Journal of Economics and Sociology* (AmEcon), *Applied Psycholinguistics* (Psycho), *Cambridge Journal of Economics* (Econ), *Cell*, *English Language and Linguistics* (Ling), *Journal of Moral Education* (Edu), *Journal of Women's History* (Hist), and *Science* (Sci). For more details see section 4.

3. The ARCHER corpus (A Representative Corpus of Historical English Registers) was compiled as part of a project on historical English registers (cf. Biber, Finegan & Atkinson 1994).

scientific argumentation, these formulae fulfill a twofold interpersonal function: On the one hand, they make the act of constructing an argument visible (through the use of a non-passivized agent-oriented verb); on the other hand, they avoid direct reference to the author who is making his or her argument, thereby minimizing politeness violations. These functions support the assumption of formulaicity being "a tool for social interaction" (Wray & Perkins 2000: 13).

2. The abstract as a genre of scientific discourse

Scientific discourse is a register in which the expression of agenthood is generally minimized (Atkinson 1992; Biber 1988; Orasan 2001). Its syntax is usually characterized by a high number of impersonal and agentless constructions, such as nominalizations (*a regression standardization revealed that…*), gerunds (*after separating the molecules…*), or short passives (*it was found that…*). The overall reason for these features is a need for objectification, which means that the individual author (who is usually reporting his or her own research) is generally not profiled. However, scientific discourse is not purely informative. Biber et al. (1999: 16), for example, characterize academic discourse as a genre whose main communicative purpose is information, argumentation and/or explanation. The persuasive function of scientific texts is also recognized in academic writing textbooks (Alley 1996; Hansen 1998; Penrose & Katz 1998), and stylistic advice is given accordingly (see 5.2).

In the course of the history of scientific discourse, argumentation has not always been the prevailing function. Bazerman (1988: 65–66) discusses how experiments moved "from any made or done thing, to an intentional investigation, to a test of a theory, to finally a proof of, or evidence for, a claim." The "focus on facts over argument" (Gross, Harmon & Reidy 2002: 19) in early scientific discourse – which was dominated by what we would now call the natural or life sciences – relied on a "confidence that the data will speak for themselves" (ibid.). In modern scientific discourse, by contrast, a finding has only the scientific significance that the scientists "bestow upon it by the power of their argument" (ibid.: 26–27). While scientists have always aimed at presenting empirical evidence, making a convincing argument for a specific analysis in the context of existing research has become increasingly important.

For most historical linguists, the rise of a scientific English begins with the foundation of the Royal Society of London in 1660; more specifically, with the establishment of the *Philosophical Transactions* five years later (Atkinson 1996; Bazerman 1988). Abstracts, however, are a more recent genre: As the data from Gross, Harmon & Reidy (2002: 132) show, abstracts were only "an occasional practice"

in 19th century articles, while regular abstracting became an established practice in the second half of the century.[4]

What goes into an abstract is determined by its context. While abstracts for conference presentations sometimes allow for a mere declaration of the intentions of the authors, the abstract that precedes a published article – the kind of abstract we focus on here – generally gives the gist of the article, which includes information about the result of the research. It "considers the article as a whole and then makes a representation of it" (Bazerman 1988: 220) and can thus be regarded as a "genre of distillation" (Swales 1990: 179). Abstracts therefore add one more layer to the long process of research, writing, and polishing that goes into turning a manuscript into a published research article (Swales 2004), and the language resulting from this process has been conclusively described as highly integrated, abstract, and impersonal (Swales 1990).[5] From the point of view of the reader, however, abstracts preceding journal articles are particularly prone to undergo "non-linear use", or skimming (Gledhill 1995: 13). Considering both the writer's as well as as the reader's needs in such a context, abstracts are rather likely to bring about formulaic language. However, formulae in such a context tend to function more as "signals of posture […] than in terms of conveying subject matter" (ibid.: 11).

Abstracts typically contain two kinds of information: (a) factual information about what has been done (why the topic is interesting, which kind of research has been carried out, which methods have been used) as well as (b) an evaluation of these facts and the direction of the argument. The first type of information is what we refer to as *reported events*; and to the second we refer as *reporting events*. Reported events are often, though not always, characterized by the use of past tense and, since we deal here with a larger variety of events, they are less likely to undergo formularization. Examples are given in (1):

4. According to Gross, Harmon & Reidy (2002: 176), in the first quarter of the 20th century only 14% of the journals studied have abstracts preceding the articles, but 81% in the third quarter and 95% of the articles in the last quarter of the 20th century. This is in line with the data for the journals that we looked at: For example, *Science* started printing abstracts in 1977; *AmEcon* printed summaries from as early as 1946 (Vol. 5), but started calling them "abstracts" much later (in 1964).

5. Many journals make abstracts freely available online and restrict access to the full article to subscribers, or offer pay-per-article access. Obviously, these practices increase the need for the abstract to be maximally informative as well as enticing. The influential APA manual reminds writers that an abstract, once published, will lead a long life as part of collections of abstracts and can thus be "the most important paragraph in your article" (APA 2001: 12).

(1) a. To test this hypothesis, **we studied** leptin-deficient and leptin receptor–deficient mice that are obese and hypogonadic. (Cell1)[6]

b. **We also investigated** the role of alphabetic skills and socioliteracy variables …. (Psycho6)

Reported events are not confined to actions carried out by the author(s) of the paper, but they may also include information about previous research (e.g., *Moral reasoning theorists… have tended to explain unethical behaviour by assuming…*, Edu1). By contrast, what we refer to here as "reporting" events (Swales 1990)[7] are the speech events that are invariably linked to the author(s) of a paper. Reporting is usually realized in present tense, sometimes in future tense, and since it involves acts of evaluation and argumentation, it is usually built around agent-oriented verbs of communication (Levin 1993), i.e., verbs like *argue, conclude, claim, consider, demonstrate, propose, reveal, show,* and *suggest*. Examples are given in (2):

(2) a. In what follows, **I posit** that moral educators can learn not only from… (Edu15)

b. **This paper argues** that, based on the ontological insights of critical realism, epistemological guidelines can be established…. (Econ11)

Due to the syntactic characteristics of reporting verbs, reporting events usually involve a bi-clausal structure with a complement clause introduced by *that*. Because of the integral function of reporting and argumentation for the purpose of an abstract, this bi-clausal pattern forms the basis of various reporting event formulae.

3. Formulaic language in the linguistic realization of scientific argumentation

While, due to a range of topics and research methods, reported events vary considerably across academic fields, the act of reporting is potentially more similar across the disciplines. In this section, we focus, first, on the use of reporting verbs in different linguistic constructions and then show how, among these, the *paper* construction shows special collocations in a specific kind of context.

6. Abstracts are identified by the journal they appeared in and by their number in our corpus; see 4.1.1 for details. Bold print in examples is all ours.

7. Swales (1990: 150) also emphasizes the role of reporting verbs in research articles, but focuses on their function of making reference to other people's research: "The RA [research article] author employs a 'reporting' verb (*show, establish, claim*, etc.) to introduce previous researchers and their findings."

3.1 Four reporting strategies

In his case study of the "Great Devonian Controversy," a 19th century dispute in the new science of geology about the correct interpretation of geological formations in Devonshire (England), Rudwick (1985) gives a detailed account of strategies for building an acceptable argument. He compares the "discovery" tradition (there are facts out there that simply need to be reported and will speak for themselves) and the "construction" tradition (the acceptance of an argument will depend on how convincingly it is made and to whom it is presented) and concludes that

> "…neither 'discovery' nor 'construction' is by itself an adequate metaphor for the production of scientific knowledge. The outcome of research is neither the unproblematic disclosure of the natural world nor a mere artifact of social negotiation. The metaphor of shaping – or, in the original sense of the term, forging – has been used… as a less inadequate image." (Rudwick 1985: 454)

Reporting events are where the shaping of an argument takes place most visibly. Even within a limited set of reporting verbs, there are a number of options for constructing these events. From the viewpoint of the verbs' semantics, the most natural choice to express a reporting event is the **agent construction**: Verbs like *argue* have two arguments, so the external argument should be realized as an agent in the position of the subject, while the internal argument is realized as a noun phrase or a complement clause, as illustrated in (3):

(3) a. In this paper **I appraise** John Wilson's ideal of (erotic) love between equals. (Edu21)

b. Simply put, **we argue** that Thorne uses faulty analysis… (AmEcon14)

Style manuals for composition and academic writing promote the agent construction as "vigorous" (APA 2001). The underlying idea is that the agent construction is maximally clear in terms of who did what to which effect. However, the principle of maximal clarity may very well be in conflict with other principles of discourse, in particular with politeness principles (Brown & Levinson 1987). Claims presented in a scientific article often imply a criticism or reformulation of a claim brought forward by somebody else and thus constitute "face threatening acts" (Garces-Conejos & Sanchez-Maccaro 1998; Myers 1985; 1989). To satisfy politeness needs, non-agentive constructions are used.[8] One of them is the **passive construction**,

8. There are of course other motivations for using the passive, such as the need for making the logical object the topic, and thus the subject, of the sentence, and the still-existing attitude that one should not use first-person pronouns in scientific writing (Alley 1996: 107), but these matter more with respect to reported events. For more details on the stylistic evaluation of the passive in reporting, see section 5.2.

as illustrated in (4). The passive expresses the same propositional content as the active, but backgrounds the agent by turning it into an implicit argument. Passives of verbs that have clausal complements usually result in impersonal constructions, as in (4):

(4) a. **It is argued** that the authority of the parent is in important respects different from the authority standardly ascribed to the teacher. (Edu21)

b. H3 hyperacetylation **is proposed** as a molecular mechanism coupling enhancer activity to accessibility for V(D)J recombination. (Sci38)

There are two more verbal strategies that are often used to express reporting events. Both combine active morphology with inanimate subjects and are, strictly speaking, in conflict with the verbs' selectional restrictions: In the ***fact* construction**, illustrated in (5), the subject position is taken by an abstract, non-agentive noun phrase like *these facts* or *these results*. This gives the impression that the facts can speak for themselves, without any intervening or interpreting agent.

(5) a. **The structural data demonstrate** how GDIs serve as negative regulators of small GTP-binding proteins and how the isoprenoid moeity is utilized in this critical regulatory interaction. (Cell13)

b. **These results contradict** the notion that metal bioavailability in sediments is controlled by geochemical equilibration of metals between porewater and reactive sulfides…. (Sci9)

We will see that this strategy is particularly popular in the experiment-based natural sciences (cf. 4.1.2) and that it is well documented historically (cf. 4.2.2): It seems to result quite naturally from scientific discourse being object-oriented, rather than "author-oriented" (Atkinson 1996: 359–360). The advantage over the passive seems to be that the active voice is preserved, while the reader's mind is nonetheless focused "on the things of the laboratory and the natural world" (Gross, Harmon & Reidy 2002: 231).

Finally, the ***paper* construction** is illustrated by the examples in (6). Rather than focusing on facts and findings, this linguistic strategy draws attention to the textual quality of the argument that is being made: The paper itself, an entity that comes into existence through the very act of arguing, advances to the subject of the reporting verb.

(6) a. **This paper argues** that, as far as theories of value and money are concerned, Marx and Menger have more in common than has been traditionally maintained. (Econ4)

b. **This article suggests** that the new institutionalism contains ambiguous and contradictory notions of change. (AmEcon10)

In the remainder of this article, we place this *paper* construction in context, explore its distribution in actual discourse and investigate what is gained by exchanging the agentive subject of reporting verbs for an inanimate, *paper*-like subject.[9]

3.2 Effects produced by the *paper* construction

Syntactically, collocations of a communicative verb and a *paper*-like subject appear to be related to the locative subject construction, such as *This room will sleep* 3 *people* (Levin 1993). But while in the locative subject construction the location argument exists prior to the event expressed by the verb, the subject in the *paper* construction is something that comes into existence through the event described by the verb. Furthermore, the locative subject construction expresses a state or quality, while the *paper* construction expresses a dynamic event, normally associated with a human agent. An article is not just the location or the scene for the argument that is presented, as in (7) (a), the argument is construed through the article, as in (7) (b).

(7) a. **In this article we present** results on the status of moral development of apprentices in the business context. (Edu2)

b. **This article analyzes** the concepts of motherhood… (Hist2)

One might think that examples like (7) (b) first and foremost have the advantage of being shorter – compact linguistic expressions are valued highly in a genre with constraints on length. However, there is really no need to express the location of the argument at all. In (7) (a), for example, one could just as well leave out the prepositional phrase *in this article*, as the positioning of the abstract makes it perfectly clear which longer text the abstract refers to. The most obvious effect of the *paper* construction is in fact that explicit reference to the "true" agent is avoided, while its presence as the writer of the paper is evoked.

In paper formulae, the paper or article advances from a pure location in which an observation is placed to the argument constructor itself. This is in line with long-term developments in the rhetoric of scientific discourse: In his account of changes in "scientific doings" in the Transactions of the Royal Society, Bazerman (1988) notes that the nature of the experiment changed from "a clear window of a self-revealing nature" to "ways of proving or supporting general claims" in conjunction with the scientific project itself developing from "individual interactions with nature" to a "communal project of constructing a world of claims" (ibid.: 78). The discovery of facts by an individual has been replaced by a detailed "process of negotiation" (Wood 2001: 75) between the author, referees, and editors, which takes place before anything gets published. The researcher is not someone who simply reports facts that will speak for themselves anymore – the researcher has to make a claim that is plausible and acceptable to a large community. In the next section we will examine how the *paper* construction is employed to achieve this effect.

9. Variations on the noun *paper* include *article, study,* and *essay* (the latter only attested in one journal, *AmEcon*).

4. Formulaic language with *paper*-like subjects in scientific English: Two corpus studies

4.1 Synchronic study

4.1.1 *Corpus information*

For our synchronic study[10] we examined 160 abstracts from 8 different scholarly journals (20 abstracts per journal), which appeared in 1999 and 2000 in the UK or in the US: *The American Journal of Economics and Sociology* (AmEcon), published by Blackwell, *Applied Psycholinguistics* (Psycho), published by Cambridge University Press, *Cambridge Journal of Economics* (Econ), published by Oxford University Press, *Cell*, published by Cell Press, *English Language and Linguistics* (Ling), published by Cambridge University Press, *Journal of Moral Education* (Edu), published by Carfax Publishing, *Journal of Women's History* (Hist), published at the time by Indiana University Press, and *Science* (Sci), published by the American Association for the Advancement of Science, with assistance of Stanford University's HighWire Press. All journals are peer-reviewed. The journals, most of which appear quarterly, were selected to represent different academic disciplines from four different areas of research (natural sciences, social sciences and economics, history and philosophy, linguistics). All abstracts directly preceded the corresponding research article. We sampled the abstracts through electronic subscriptions held by the University of Wisconsin-Madison and the University of Düsseldorf. None of the journals provided formal instructions on how to write an abstract for a specific journal, with the exception of constraints on length, but these limits were not always respected.

Table 1. Length of abstracts (not counting abstract title)

	Sci	Econ	Cell	AmEcon	Hist	Edu	Ling	Psycho
Words per 20 abstracts	1924	2124	2433	2890	2899	2930	3194	3410
Mean length of abstract	96.2	106.2	121.7	144.5	145	146.5	159.7	170.5

10. Corpus information and more detailed results on the patterns of agentivity in abstracts are presented in Dorgeloh & Wanner (2003).

We compared the number of reporting events per abstract among journals and then examined the linguistic forms that were used to realize them.[11]

4.1.2 *Reporting across academic disciplines*

Since *paper* constructions are a specific way to express a reporting event, we first looked at the number of reporting events across the disciplines. Typical reporting verbs used to express reporting events are *argue*, *conclude*, *demonstrate, indicate, present, recommend, show,* and *suggest*.[12] Overall, we found that abstracts without reporting events are very rare (Dorgeloh & Wanner 2003); they occurred predominantly in two journals, *Science* and *Hist*. There was also some intrajournal variation, again especially in *Hist*, but across journals some generalizations regarding the distribution of reporting strategies can clearly be made.

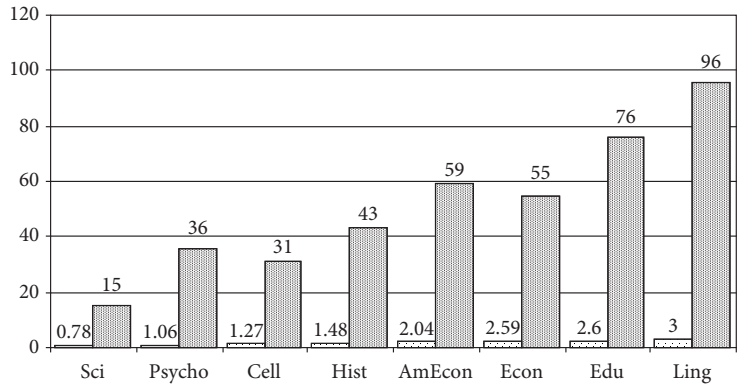


Figure 1. Number of reporting events (tokens per 100 words and absolute number of tokens in 20 abstracts).

11. We are aware that in all likelihood a considerable number of the abstracts has not been written by native speakers. For example, Wood (2001) estimates that about 45% of research articles in an international journal like *Science* are written by non-native speakers of English. We will not explore this issue any further, though, and will simply assume that all published articles and abstracts conform to an acceptable standard of international English (cf. Crystal 1998).

12. According to Biber et al. (1999: 668) the most common verbs to occur with *that*-clauses in academic prose are *suggest* and *show*. These were also the most common reporting verbs in our abstract corpus.

Figure 1 shows the absolute and relative number of reporting events in each of the eight journals. The absolute and relative frequency of reporting events is lowest in *Science* and the experiment-based journals *Cell* and *Psycho* (0.78 to 1.27 reporting events per 100 words). *Econ, Edu,* and *Ling*, three journals from the humanities or social sciences, appear at the other end of the spectrum, with 2.59 to 3.0 reporting events per 100 words. *Hist* takes an intermediate position: Unlike any other journal from the humanities or social sciences it occasionally has abstracts without any reporting event at all; discarding these, abstracts in *Hist* are more like the ones in other social sciences (*Econ and AmEcon*) on the reporting scale.

The findings presented in Figure 1 can be summed up as follows. It seems that, in the tradition of the "discovery" approach to gaining knowledge (Rudwick 1985), the interpreter role of the researcher is not as prominent in the natural sciences as it is in the humanities and social sciences. For the latter, the construction of an argument (through reporting events) takes up a considerable portion of the abstract; probably as a result of this, reporting events in these abstracts also show more linguistic variation (see 4.1.3). This difference is underlined by the very low number of reporting events in *Science*, which cannot be attributed to space restrictions only. Clearly, abstracts in *Science* are shorter than most other abstracts, many have fewer than 100 words, some have fewer than 50. This voluntary restriction to the bare-bones essentials of scientific work makes it obvious that no need is felt to focus on the author as someone who is shaping an argument. In the next section, we will further discuss the function of reporting events in different academic disciplines and see how they are reflected by the typical collocations used to realize them.

4.1.3 *Formulaic language across the disciplines*

Let us take a closer look at two of our journals which illustrate the differences between the "discovery" approach and the "construction" approach to presenting arguments in an exemplary way. As Myers (1992) has argued, reporting in the natural sciences is normally a true reporting of new facts, while in the humanities the new material to be reported can be just as well a new reading of facts that may already have been known. In *Edu*, for example, articles deal with questions like how to explain unethical behavior or what should be considered the role of poetry for human growth; they are usually not based on experiments and often deal with new answers to questions that have been asked before. In addition, articles in this journal often have a single author. Articles in *Cell*, on the other hand, discuss topics like the molecular structure of prions and genetic sequencing and almost always involve some kind of experimental work, carried out by a team of researchers. Abstracts in *Cell* are about 20% shorter than in *Edu* (121.7 vs. 146.5 words), but the differences in the use of reporting events is even more striking: For *Cell* the overall number is 31 (1.27 reporting events per 100 words), for *Edu* it is more than twice as high, both in absolute

and relative numbers (76 reporting events altogether, 2.6 per 100 words). As to the linguistic realization of these events, surprisingly, neither the active nor the passive construction is the most common form of reporting in these abstracts: The *paper* construction is most common in *Edu* (36.8%) and the least common construction in *Cell* (6.5%), while the complementary picture arises for the construction with *fact*-like subjects, which is the favored one in *Cell* (45.2%) and by far the least preferred strategy in *Edu* (5.3%). This pattern goes beyond these two journals, as illustrated in Figure 2.[13] In four out of the six journals considered (*Econ*, *AmEcon*, *Edu*, and *Ling)*, a relatively high number of reporting events correlates with a high frequency of *paper* constructions. In the more experiment-based abstracts, such as *Cell* and *Psycho*, the *paper* construction is less frequent. Figure 2 also shows the popularity of formulaic expressions with *fact*-like subjects for experiment-based publications: Not only is it the strategy with the highest number of tokens in both *Cell* and *Psycho*, but the absolute number of tokens of *fact* constructions is higher in *Cell* and *Psycho* than in any of the other journals. This may seem surprising, considering the fact that *Ling* has three times as many reporting events as *Cell*.

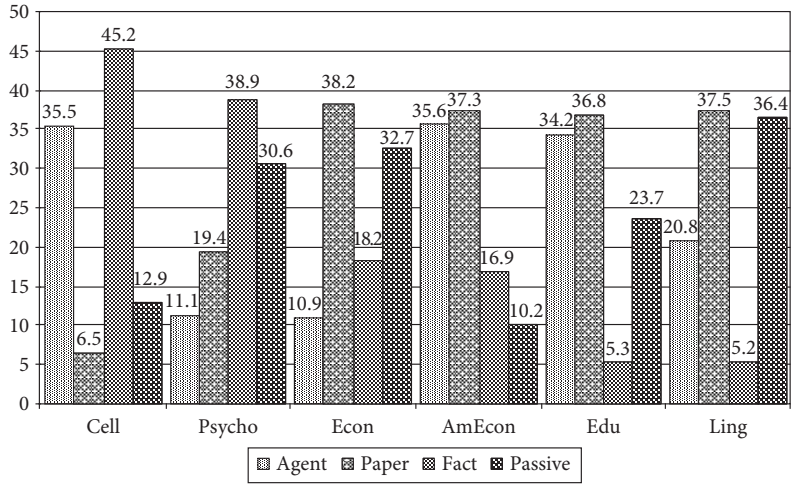


Figure 2. Overview – reporting strategies in six selected journals (in percent of reporting events).

13. We focus on 6 of the 8 journals here. We leave out *Science*, which is similar to *Cell*, but has much shorter abstracts than any of the other journals, and *Hist*, because of its substantial intrajournal variation.

In summary, the data from our abstract corpus indicate a correlation between the role of reporting and the occurrence of formulae with *paper*-like subjects. In particular, they show that the *paper* construction is the preferred option in journals with an overall high proportion of reporting events (*Econ, AmEcon, Edu, Ling*), which is more characteristic of journals from the humanities.[14] By contrast, disciplines that are more focused on experiment-based findings do not necessarily have abstracts that are shorter, but they have fewer reporting events and also less linguistic variation within them. Although ordinary actives and passives seem to be the most natural choices for expressing reporting events from the viewpoint of lexical semantics, neither of them is the most frequent strategy in any journal in our corpus. While journals with a high proportion of reporting events also have the highest proportion of *paper*-based formulae, journals with a lower frequency of reporting events have *fact*-like subjects as the most frequent reporting strategy. None of the journals with comparably little reporting has more argumentation built on the *paper* strategy than on the *fact* strategy. Considering the attention that the choice between active and passive gets in style manuals (the active as a desirable strategy, the passive as a construction that is considered overused), these results are quite surprising and highlight the flexibility to fill the subject position with non-agent arguments that English grammar is known for (Levin 1993).

4.2 Diachronic study

From a synchronic viewpoint, we found that the *paper* construction is used more frequently in contexts in which reporting overall figures prominently. Our interpretation of these data is that historically the rise of this pattern is contingent on the function of explicit reporting, which is in turn constitutive of the modern scientific article and its condensed version, the abstract. In order to explore this assumption, we now present data from a study of a historic corpus of scientific English.

4.2.1 *Corpus information*

For historical data on reporting in scientific English we used the ARCHER corpus. We analyzed its category of *Science*, which consists of seven subcorpora from 1650 to 1990 of about 20,000 words each. Text samples in this corpus come from longer articles and monographs, which is why reporting events are, in absolute terms,

14. We use the labels "humanities" and "social sciences" only with respect to the abstracts from our corpus. There are of course different traditions of subjects and approaches in any academic field. There is a strong tradition of empirical, survey-based work in the social sciences, for example, but predominantly the abstracts that we sampled did not belong to this tradition.

much less frequent than in the abstract corpus. The texts are all from the *Philosophical Transactions* of the Royal Society.[15]

We investigated the use of a limited set of 8 verbs, which are commonly used in the abstract corpus to express reporting events: Five of them represent typical speech act verbs of scientific reporting proper (*argue, demonstrate, indicate, show, suggest*), while the other three originally refer to scientific work (*examine, explore, find*), but in this context are also frequently used to express mental activities (*We examine/explore/find that…*).

4.2.2 *Reporting constructions in historical scientific English*

The use of these eight reporting verbs in the ARCHER category *Science* is presented in Figure 3. These results include the occurrence of reporting events proper (in present tense), but also past tense forms, because, in genres other than abstracts, these often refer to an act of reporting (*Findings suggested that…*). The chart shows a steady increase of these reporting verbs over time, which reflects the growing role of reporting as such.

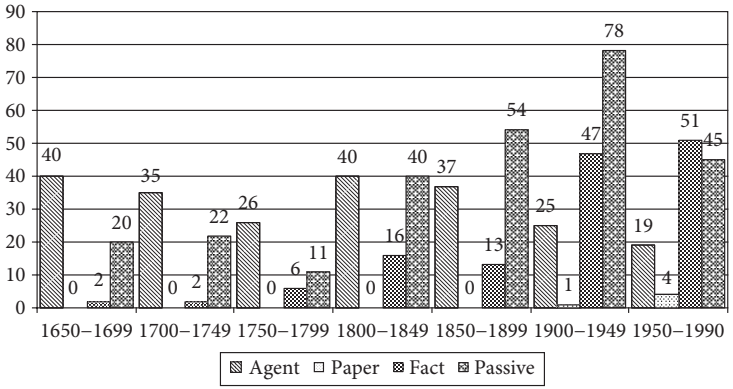


Figure 3. Realization of selected reporting verbs (*argue, demonstrate, examine, explore, find, indicate, show, suggest*) in ARCHER (tokens per 20000 words).

If we look at the distribution of the four verbal strategies in more detail, we can see that the data confirm a widespread generalization about the language of

15. The Royal Society was founded in 1660 as a body that met weekly to witness and discuss experiments. At the time of its foundation there was no comparable institution or journal for the humanities, i.e., the samples in our corpus are all more or less from the natural sciences.

science: its preference for using the passive. There is an increase of passive forms until the penultimate time segment, and the passive is used more frequently than the agentive active.

More interestingly, however, there is also a significant development among the three active alternatives: While the agent construction decreases over time, but never appears to become altogether unusual, it is the fact construction that rises steadily, from an almost negligible frequency up to the end of 18th century (2–6 tokens for 8 selected verbs) to a more frequent usage in the 19th (13–16 tokens) and with a sharp rise in the 20th century (47–51 tokens). And it is only then that we also find instances of the *paper* construction to a notable extent for the first time. The rise of the *paper* construction thus follows the establishment of the *fact* construction as one of the accepted (by virtue of being published) forms for scientific reporting (in the last stage the *fact* construction is even slightly more frequent than the passive), and it runs parallel to the emergence of the scientific abstract as a genre (cf. section 2). Of course, the proportion of the passive remains quite substantial at all stages, but one has to bear in mind that the scientific text samples in ARCHER come from longer articles and monographs, which do not constitute a discourse consisting of explicit, institutionalized reporting.

The *paper* construction thus seems to be a very modern and a genre-specific phenomenon, and its emergence cannot be traced back too far in the history of scientific English. However, *paper*-based formulae have a close relative in the usage of reporting verbs with other inanimate subjects, i.e., in the *fact* construction. We suggest that the increasing use of combinations of the type *results suggest…* in scientific English prepared the way for subjects referring to the publication itself – a development that was in turn plausibly triggered by an increasing number of relevant publications, notably journals, in the second half of the 20th century.

5. Discussion of results

5.1 Politeness concerns

As we have pointed out elsewhere (Dorgeloh & Wanner 2004), reporting strategies in scientific discourse have also been accounted for in the light of politeness considerations. Within this framework (Brown & Levinson 1978), non-agentive, *paper*- as well as *fact*-like subjects are said to result from a need for negative politeness: Both strategies assure the readers "that the writers do not intend to infringe on their wants" (Myers 1989: 12). The writers "need to manifest deference to and solidarity with their respective research communities, which, through their various gatekeeping roles, exert considerable power and influence" (Swales 2004: 218).

Negative politeness is achieved by combining an impersonalized subject (*this finding*, *this paper*) with the hedging provided by a reporting matrix verb (*someone or something suggests that a is b* rather than just *a is b*). Following Myers, the result is that scientific claims receive "some impersonal agency" (1989: 17).

The data presented in section 4 confirm this position and they also emphasize the role of reporting as such for the emergence of formulae: As shown in 4.1.2, reporting events are particularly prominent in the humanities, i.e., in the disciplines in which a finding often involves a re-evaluation of facts or beliefs. In 4.1.3 we argued that this explains the high frequency of the *paper* construction in abstracts from disciplines that are not generally experiment-based. In these areas of research, the article is not just the location where a claim is made and an argument is established, it is the article itself that is making this claim. This strategy seems to meet negative politeness requirements best: The authors of conflicting claims only serve their cause by writing it down – they appear to be advocates of ideas rather than defendants under personal scrutiny. The abstracts from the natural sciences still focus more strongly on data and results and trust these will speak for themselves, which is reflected in the preference for the *fact* construction.

5.2 Changes in the rhetoric of scientific discourse

Swales (1990) characterizes the academic community as a discourse community that has a broadly agreed-upon set of goals, mechanisms of intercommunication, and specific genres and registers of communication. Bazerman (1988: 155) emphasizes that "the forms of scientific representation emerged simultaneously and dialectically with the activity of science and the social structure of the scientific community." Changes in the structure of the community have led to changes in the organization and representation of scientific discourse. We consider the growing popularity of the *paper* construction to be one of the results of those changes (rather than a remnant of the "discovery" tradition).

More than any other reporting strategy, the *paper* construction depends on the recognition of the role of the research article. Obviously, there would be no *paper*-like subjects without the rise of the research article, or *paper*, and the orientation of professional scientific discourse "from book to article science" (Bazerman 1988: 81). The research article has become the prime instrument in presenting new material that has already undergone some validation (through peer reviews). The growing significance of the research article is related to the increase in the number of scholarly journals that members of the scientific community have to keep up with. As a response to these challenges, printing abstracts of articles has become standard practice. Readers of research articles are "extremely fickle" (Swales 1990: 179) – some will read only the title of an article and decide not to read

the whole article, others will at least advance to the abstract to find out whether or not the article will be of interest to them. The point of an abstract is to give a representation of the argument that will be made; therefore, reporting events are quite prominent in abstracts, and preferences for certain reporting strategies will be noticeable in abstracts.

As laid out in section 2, scientific discourse has undergone rhetorical shifts, from narrative to presentation of facts to the negotiation of assumptions and conclusions (Atkinson 1992; Myers 1992). These changes in the development of scientific discourse go beyond the emergence of new genres. With reference to Barthes (1975), Swales (1990: 112) characterizes the rhetoric of the classic research article as one in which "[t]he art of the matter, as far as the creation of facts is concerned, lies in deceiving the reader into thinking that there is no rhetoric, that research reporting is indeed 'writing degree zero' … and that the facts are indeed speaking for themselves." This style – "authors adopt humility before the facts" (Bazerman 1988: 142) – is best represented by the *fact* construction, which, as shown in Figure 2, is the preferred reporting strategy in *Cell* and *Psycho* (also in *Science*, not included in Figure 2), i.e., in journals positioned firmly within the experiment-based tradition.

Paper-based formulae, on the other hand, are a rather recent phenomenon (cf. 4.2.2) and reflect a different direction in what is considered report-worthy material. They coincide with a constructional approach to science, in which the role of the scientist in creating data and the role of the author in presenting and interpreting them is stressed. Rudwick (1985: 453) considers analyses according to which accounts of scientific knowledge are "virtually *nothing but* a social construction" as extreme, though he concedes that they have "some plausibility" when they are based on "forms of modern research in which the 'raw' natural world is almost excluded" (italics in original). He refers to something like the use of elaborate instrumentation or purified chemicals, but if we take his assessment beyond the context of just one discipline, it also applies to the contrast between the humanities and the natural sciences. Research articles in the humanities do not deal with the raw natural world and it is no surprise that they make more use of linguistic strategies that reflect the nature of an argument as something being constructed via language, in particular of the *paper* construction.

Our synchronic data, as presented in Figure 2, show that there are five journals in which the *paper* construction is the most preferred form of reporting in abstracts (*Ling*, *Edu*, *AmEcon*, *Hist*, and *Econ*). Most of these journals are also characterized by a rather high frequency of reporting events as such, which – in principle – could also have been realized by any of the three alternative strategies. Though a considerable number of articles in these journals is based on empirical work, most are not written in the tradition of the natural sciences, in which the

laboratory experience is the main source of the scholarly argument. The point of the article is not to enable the reader to become a virtual witness of an experiment or to duplicate it; the point is to present and make visible the construction of an argument that is acceptable to the research community.[16]

The success of *paper*-based formulae is also related to the stigmatization of the passive and, to some extent, collocations with *fact*-like subjects. According to Baron (1989: 19) "by the 1940s the passive, with its deletable agent… became associated not simply with the mildly distasteful traits of wordiness and confusion, but with the even more negative practice of conscious deception by deliberately hiding the doer of the action." Modern textbooks on scientific writing strongly advise students against the use of the passive, which is perceived as a construction that "saps … energy and leaves dead words on the page" (Alley 1996: 106). It is also considered the expression of a "detached persona," conveying the impression "that knowledge is self-existent, an object that is discovered rather than at least partially created through researchers' methods of collecting and interpreting evidence" (Hansen 1998: 437). Hansen points out that inanimate subjects like study (in *this study investigates…*) create the same – and undesirable – "detached persona" in writing. Style manuals also warn against the use of the *fact* strategy because of its perceived anthropomorphism (APA 2001: 38).

How, then, do we explain the success of the *paper* construction? Based on our analysis of historical data in 4.2, we suggested that this specific formula is a successor of the *fact* construction rather than a version of it. Although the agent is not named explicitly, the *paper*-subject draws attention to his or her work as a writer, interpreter and, to use Rudwick's and Bazerman's metaphor, "shaper" of knowledge. Unlike the *fact* construction, the *paper* construction does not give the impression that "that knowledge is objective, impersonal, and disembodied, existing apart from people who know it" (Hansen 1998: 438). While the data presented in Figure 3 show an increasing inclination for using the passive in the first three and a half centuries of scientific English, this trend appears to have stopped in the late 20th century. Instead, since the beginning of that century, the *fact* construction has evolved, especially in the natural sciences, possibly preparing

16. Differences in the significance of active reporting are reflected in style manuals from different disciplines. The APA [American Psychological Association] Manual, the most influential style manual for the social sciences, devotes about 4 pages to the subject of how to write an abstract and gives very clear stylistic advice, including a warning against the fact strategy (APA 2001). By contrast, the CBE [Council of Biology Editors] Manual, one of the leading manuals for the life sciences, limits itself to short technical advice on the subject of abstract-writing (CBE 1994).

the way for the *paper* construction in particular genres (such as abstracts) and disciplines (such as the humanities).

6. Conclusion

We have analyzed the emergence and the use of the *paper* construction (*This paper argues…*), which gives rise to specific, widely used formulae that express a reporting event, i.e., an explicit act of scientific argumentation. Syntactically, these are constructions in which an adverbial (the location or instrument of an event) advances to the subject position, thereby allowing for the non-expression of the agent without having to resort to the passive, which is often stigmatized as impersonal and pseudo-objective (Baron 1989; Leunen 1978). Pragmatically, the *paper* construction is in line with the tradition of minimizing face-threatening acts and hedging the force of the argument in scientific discourse (Myers 1989). Crucially, it also stresses the role of the article or paper for the construction of the argument – a role that has increased substantially over time.

In a synchronic study, we focused on abstracts of research articles, which have a high proportion of reporting events. Due to the semantics of the verbs involved, these are events in which an individual author should be most visible. On the other hand, these are also events with a high potential for politeness violations. The need to tone down the face-threatening force of the claims that are made does not seem to be served best by using the passive. Instead, we found a high proportion of the *paper* construction in the humanities, while in abstracts of experiment-based articles there was a related preference, but it was for what we termed the *fact* construction (*These data suggest…*). We could generalize that, with the exception of one journal, the higher the number of reporting events is, the higher the proportion (not just the absolute number of tokens) of the *paper* construction is likely to be. Across the board, both constructions were more frequent than the agentive active (*We suggest…*) and the morphologically marked passive (*It is suggested…*).

It has been observed that genres tend to regularize and conventionalize language since they "increase the likelihood of successful, forceful communication" (Bazerman 1988: 23). The genre of the research article abstract is relatively new, but well established through the conventions of modern publishing. It functions like a distilled version of the research article but does not go through all the corresponding rhetorical moves. From our diachronic study, we saw that the *paper* construction is about as modern as the genre abstract itself, and that it is historically preceded by the *fact* construction, which has steadily increased in usage and has become even more frequent

than the passive. On the basis of our findings we concluded that the *fact* construction paved the way for the *paper* construction, as it is also a construction that combines an agent-oriented verb with an inanimate subject. More so than the *paper* construction, the *fact* construction presents scientific arguments as self-contained and non-constructional in nature.

Our study is obviously limited with respect to the size of the abstract corpus. Another limitation is that we focused on verbal reporting strategies only. There are a number of other ways to realize reporting, including nominalization (*Our central proposition is that…*)[17] or adjectives (*It is clear that…*). We also did not look at stylistic factors, such as attempts at variation of strategies, or at the model character of abstracts published in a specific journal. For instance, 7 out of 20 abstracts from *Ling* in our corpus begin with the formulaic subject *This paper,* and 9 out of 20 abstracts in *Hist* with *This article*. It is not far-fetched to assume that some writers model their abstracts after published abstracts in the target journal.

We conclude (a) that behind the *paper*-based formulae that we discussed here stands a productive strategy of reporting, (b) that these specific formulae are a product of the dominant role of the research article (and of the growing role of abstracts that precede and summarize them), and (c) that they reflect some general changes in the rhetoric of scientific discourse. They allow the writer to deprofile the agent without resorting to the stigmatized passive. This is something the *paper* construction shares with the *fact* construction, but in contrast to the latter, the *paper* construction does not present an argument as neutral and self-evident, but reminds the reader of the constructional (and textual) nature of the argument. Although there is a general trend to express the involvement of the writer, at present this seems to be more of a concern in the humanities and social sciences. This difference explains the preference for collocations with metatextual subjects of the *paper*-kind in journals in the humanities – and the (lasting) popularity of experiment-based, *fact*-like subjects in the natural sciences.

References

Alley, Michael. 1996. *The craft of scientific writing*. Berlin: Springer.

APA (ed.) 2001. *Publication manual.* Washington DC: American Psychological Association.

Atkinson, Dwight. 1992. The evolution of medical research writing from 1735 to 1985. *Applied Linguistics* 13: 337–374.

Atkinson, Dwight. 1996. The philosophical transactions of the royal society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society* 25: 333–371.

17. See Kerz & Haas (this volume) on this kind of construction.

Baron, Dennis. 1989. *Declining grammar*. Urbana IL: National Council of Teachers of English.

Barthes, Roland. 1975. *The pleasure of the text*. New York NY: Hill.

Bazerman, Charles. 1988. *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison WI: The University of Wisconsin Press.

Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge: CUP.

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

Brown, Penelope & Stephen Levinson. 1978. Politeness in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, Esther Goody (ed.) 56–289. Cambridge: CUP.

Brown, Penelope & Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: CUP.

CBE (ed.) 1994. *Scientific style and format. The CBE [Council of Biology Editors] manual for authors, editors, and publishers.* Cambridge: CUP.

Cowie, Anthonie P. 1992. Multiword lexical units and communicative language teaching. In *Vocabulary and applied linguistics*, P. Arnaud & H. Béjoint (Eds) 1–12. Basingstoke: Macmillan.

Crystal, David. 1989. *English as a global language*. Cambridge: CUP.

Dorgeloh, Heidrun & Anja Wanner. 2003. Too abstract for agents? The syntax and semantics of agentivity in English research articles. In *Mediating between concepts and grammar*, H. Härtl & H. Tappe (Eds) 433–453. Berlin: Mouton de Gruyter.

Dorgeloh, Heidrun & Anja Wanner. 2004. The limits of variation in scientific abstracts: Syntactic and functional constraints. *ITL Review of Applied Linguistics* 143–144: 37–60.

Garces-Conejos, Pilar & Antonia Sanchez-Maccaro. 1998. Scientific discourse as interaction: scientific articles vs. popularisations. In *Linguistic choice across genres: variation in spoken and written English*, Antonia Sanchez-Maccaro & Ronald Carter (Eds) 173–190. Amsterdam: John Benjamins.

Gledhill, Chris. 1995. Collocation and genre analysis. Zeitschrift für Anglistik und Amerikanistik 43: 11–36.

Gledhill, Chris. 2000. *Collocations in science writing.* Tübingen: Narr.

Gross, Alan, Joseph Harmon & Michael Reidy. 2002. *Communicating science: The scientific article from the 17th century to the present*. Oxford: OUP.

Hansen, Kristine. 1998. *A rhetoric for the social sciences. A guide to academic and professional communication*. Upper Saddle River NJ: Prentice Hall.

Leunen, Mary-Claire. 1978. *A handbook for scholars*. New York NY: Alfred Knopf.

Levin, Beth. 1993. *English verb classes and alternations. A preliminary investigation*. Chicago IL: The University of Chicago Press.

Miller, Carolyn. 1984. Genre as social action. *Quarterly Journal of Speech* 70: 151–167.

Myers, Greg. 1985. Text as knowledge claims: The social construction of two biology articles. *Social Studies of Science* 15: 593–630.

Myers, Greg. 1989. The pragmatics of politeness in scientific articles. *Applied Linguistics* 10: 1–35.

Myers, Greg. 1992. "In this paper we report": speech acts and scientific facts. *Journal of Pragmatics* 17: 295–313.

Orasan, Constantin. 2001. Patterns in scientific abstracts. In *Proceedings of Corpus Linguistics* 2001 [Technical Papers 13], P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (Eds) 433–442. Lancaster: University Centre for Computer Corpus Research on Language.

Penrose, Ann & Steven Katz. 1998. *Writing in the sciences. Exploring conventions of scientific discourse*. New York NY: St. Martin's Press.
Rudwick, Martin. 1985. *The great Devonian controversy. The shaping of scientific knowledge among gentlemanly specialists*. Chicago IL: University of Chicago Press.
Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: CUP.
Swales, John. 2004. *Research genres: Explorations and applications*. Cambridge: CUP.
Wood, Alistair. 2001. International scientific English: the language of research scientists around the world. In *Research perspectives on English for academic purposes*, J. Flowerdew & M. Peacock (Eds) 71–83. Cambridge: CUP.
Wray, Alison & Michael Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20: 1–28.

Accepting responsibility at defendants' sentencing hearings

No formulas for success

M. Catherine Gruber

Abstract

This paper explores intersections between the socio-interactional functions of formulaic and non-formulaic language and the stigmatized institutional role identity of criminal defendant. Two different means through which defendants articulate acceptance of responsibility during allocution at sentencing are examined: one is formulaic due to its performative syntactic frame and limited set of lexemes; the other lacks formulaicity. Consistent with work that has portrayed allocution as fraught with pitfalls for defendants, both kinds of statements advance defendants' communicative goals in some respects, but undermine them in others. Focusing on a communicative task that privileges the affective aspect of communication instead of the (more commonly examined) social or effective aspect adds to our understanding of the diverse socio-interactional functions of formulaic language.

1. Introduction

This paper explores some of the intersections between the socio-interactional functions of formulaic and less formulaic language and institutional role identity – specifically the stigmatized institutional role identity of criminal

defendant. I focus on two different ways in which defendants articulate a claim of responsibility for their actions during federal sentencing hearings – one that is relatively more formulaic due to the way in which it exhibits a performative syntactic frame and limited set of lexemes, and one that is relatively less formulaic. Building on Gruber (in press), which articulates some of the contextual constraints that limit the effectiveness with which defendants can speak on their own behalf, this paper proposes that while formulaic statements which communicate a defendant's acceptance of responsibility may serve a defendant's interests in some respects, they appear to undermine other, perhaps more important, communicative goals.

On the one hand, the use of formulaic language is consistent with the ritual aspects of sentencing hearings; further, the formulaic features of the constructions that defendants use have clear ties to Austinian performatives, adding connotations of agentivity to defendants' statements. On the other hand, in addition to problems of ambiguity associated with the particular formula that is used, defendants who use formulas to communicate their acceptance of responsibility risk foregrounding their similarities to other defendants who have produced the same formulas. As a result of this strategy, defendants diminish their capacity to present themselves in an individualized manner. Furthermore, due to the fact that an official sentencing deduction is referred to as "acceptance of responsibility," the use of formulas containing variants of the words "accept responsibility" functions to foreground a defendant's self-interest in the upcoming sentence. This presumed self-interest potentially undermines the sincerity of a defendant's statement of responsibility. While the use of more creative language to "accept responsibility" avoids the pitfalls associated with formulaic language and appears to provide other benefits as well, it also carries potential risks. Thus, while research in related genres suggests that there exists a strong connection between creativity and perceived sincerity in the production of expressive speech acts in American culture, speaking in more creative ways may index a speaker persona that clashes with notions of speakerhood that are permitted to those who inhabit stigmatized institutional identities.

Following Pawley (2007), the approach to the study of formulaic language adopted in this paper follows in the footsteps of philosophers (e.g., Grice 1957; Austin 1962), sociologists (e.g., Goffman 1971; Tavuchis 1991), and sociolinguists (e.g., Bach & Harnish 1979; Coulmas 1981; Johnstone 1991; and Aijmer 1996, among many others) who have studied the socio-interactional functions of formulaic language. By focusing on a communicative task that privileges the affective aspect of communication instead of the (more commonly examined) social or effective aspect, it is hoped that this paper will add to understandings of the ways in which the use of formulaic and less formulaic language

provides different kinds of opportunities for meaning-making in different kinds of communicative contexts.

2. Plan for the paper

The plan for this paper is as follows: first I introduce allocution and present a brief overview of its history in U.S. District Courts. I then describe the study and highlight some of the features of the sentencing hearings, including the defendants and their offenses, at which the data were collected. Because this paper focuses on one type of defendant utterance, a brief overview of the coding system as a whole is presented in order to contextually situate the utterances of focus for this paper. Turning next to federal sentencing hearings, I review the Federal Sentencing Guidelines' sentencing grid which governs the majority of federal sentences, paying special attention to the sentencing deduction for "acceptance of responsibility." With this background in place, the paper then turns to the set of defendant statements that develop topics related to the offense and utterances that have been coded as explicitly expressing acceptance of responsibility (a/r) for the offense are explored in detail. The a/r utterances have been divided into two groups: those that are relatively more formulaic and those that lack formulaicity. The pros and cons associated with these different choices are examined via an assessment of how these different strategies both advance and undermine defendants' presumed communicative goals at sentencing hearings. Following Sadock (1994), three different aspects of language use – informational, effective, and affective – are introduced and this paper argues that discussions of formulaic language in the literature have frequently focused on communicative exchanges that privilege social or effective kinds of functions – the language used to "get things done" in the world. In contrast, apologetic allocutions at sentencing are shown to provide an opportunity for the examination of language use in which the affective or emotional aspect of communication is privileged. The different socio-interactional meanings associated with the use of formulaic vs. non-formulaic language in these different contexts suggests that differentiating speech situations in this way could lead to fruitful congruences of the meanings of formulaic and non-formulaic language in contexts that exhibit similar features.

3. Background on the right of allocution

The right of allocution is defined by the legal encyclopedia, *American Jurisprudence*, as "the early common law practice of asking the defendant whether he or she has anything to say why sentence should not be pronounced against him or her"

1998: 387). This right, although not constitutionally guaranteed, is required by the Federal Rules of Criminal Procedure.[1] The right of allocution was added as Rule 32(a)(1) in a 1966 amendment on the basis of the 1961 case of *Green v. United States*, where the court held that as a matter of good judicial administration, trial judges should unambiguously address themselves to the defendant "and leave no room for doubt that the defendant has been issued a personal invitation to speak prior to sentencing." (Wright, King, and Klein 2004: §525) Allocution is viewed primarily as a benefit for defendants because it provides the defendant the opportunity "to make a statement in his own behalf" and "to present any information in mitigation of punishment." (Wright, King, and Klein 2004: 152) Allocution also comes with risks, however. Well-known risks concern the possibility that the defendant could speak in such a way that would suggest that she had not fully accepted responsibility for the crime (O'Hear 1997; Natapoff 2005). Gruber (in press) adds to this picture by articulating a number of ways in which the context of the sentencing hearing, including its discursive constraints, functions to limit the kinds of things that defendants can say on their own behalf and undermines the effectiveness with which they can speak.

4. Data

This paper builds on the findings of my dissertation, which examined the discursive patterns of 52 apology narratives performed by defendants during allocution at sentencing hearings in federal court. The data for this paper were collected between November 2004 and March 2006 in three U.S District courtrooms which are referred to here as the courtrooms of Judge X, Judge Y, and Judge Z.[2] Judge X was a Caucasian woman; Judges Y and Z were Caucasian men. Seventeen allocutions were collected from Judge X's and Judge Y's courtrooms and 18 allocutions were collected from Judge Z's courtroom. I attended and took notes on each of the sentencing hearings and afterwards used recordings of the hearings to make my own transcripts for analysis. The 52 defendants consisted of 41 men and 11women whose ages ranged from 20 to 56, with a median age of 30. In terms of race, the dataset consists of 26 Caucasian defendants, 16 African-American defendants, 8 Native American defendants, and 1 Hispanic and 1 Asian defendant. Forty-eight of the 52 defendants had pleaded guilty

1. In the U.S. as a whole, the right of allocution is recognized in more than half of the American jurisdictions (*McGautha v. California*, 402 U.S. 183 (1971)).

2. I am very grateful for a dissertation improvement grant from the National Science Foundation's Law and Social Science Program and to the University of Chicago's Language Laboratories and Archives for the use of recording equipment.

and four had been convicted by a jury. (This rate of 92% of guilty pleas is lower than the national average of 96%.[3]) Drug and firearm-related offenses accounted for 65% of the crimes for which the defendants in this dataset were being sentenced; various types of fraud, theft, and bank robberies accounted for 25% of defendants' crimes.[4]

The statements pertaining to acceptance of responsibility constituted just one of the topics that defendants developed during their allocutions. In order to capture the range of topics and themes in defendants' allocutions, I developed a coding system consisting of eight basic codes and 26 subcodes. The eight basic codes identified statements in which defendants (A) accepted the opportunity to address the court, sometimes with *thank you* and/or polite terms of address for the judge; (B) criticized their actions or offered a personal reaction to the offense, often by means of some sort of apology; (C) talked about how they had changed or how they had learned a lesson from the experience; (D) offered information in mitigation of a harsh sentence, such as by making mention of positive things they had done or by referring to a difficult childhood or to children who needed them at home; (E) thanked family and friends for their support; (F) made reference to the sentence that was about to be imposed – sometimes requesting leniency, other times conveying their acceptance of whatever sentence would be imposed; (G) ended their turn at talk, sometimes with *thank you* or with polite terms of address for the judge; and (H) broke the frame of allocution by making reference to the physical context of the sentencing hearing, such as by asking permission to stand or referring to the microphone. The majority of defendants' allocutions employed between three and five of these eight elements. The apology narratives ranged from 4 seconds to 186 seconds in length of time that they occupied and the median allocution was about 30 seconds long. For a sample allocution, see Appendix A. (For a complete discussion of all of the codes and their distribution patterns, see Gruber 2007.)

5. Federal sentencing hearings and "acceptance of responsibility"

At federal sentencing hearings, the judge generally imposes a sentence based on the Federal Sentencing Guidelines, which link ranges of months of imprisonment with 43 offense levels and six criminal history categories. The Federal Sentencing Table is

3. According to U.S. government statistics from October 1, 2004 – September 30, 2005, 96% of all convictions of federal defendants resulted from guilty pleas. See http://www.uscourts.gov/judbus2005/appendices/d7.pdf.

4. The remaining 10% of offenses related to child pornography, assaulting a guard at a Federal Correctional Institution, escaping from a correctional institution, and using an explosive device (http://www.uscourts.gov/judbus2005/appendices/d2.pdf).

reproduced in Appendix B. For sentencing ranges that exceed 24 months, the maximum of a sentencing range does not exceed the minimum by more than 25 percent. In January 2005, in the case of *United States v. Booker*, the Supreme Court held that the Federal Sentencing Guidelines were no longer mandatory; as a result, a judge now has greater leeway to impose a sentence that is below the guidelines when s/he feels such a sentence is warranted.[5] Based upon the 2005 Sourcebook of Federal Sentencing Statistics, however, the Guidelines appear to be used in much the same way that they were before *Booker*.[6] In support of this claim, on February 20, 2007, Linda Greenhouse noted in the *New York Times* that a study by the United States Sentencing Commission last year reports that "judges have continued to impose sentences within the guidelines… in 86% of all sentences." In most cases then, it is that 25% of the lower end of the range that the judge exercises her discretion over, and on which a defendant's allocution could potentially have an effect.

Defendants get assigned to a particular range of months based on the intersection of their total offense level and their criminal history category.[7] A defendant's total offense level is determined by the base offense level associated with his/her particular crime (for example, the base offense level for robbery is 20); the base offense level is then increased and/or decreased based on the presence of additional aggravating or mitigating factors. Committing a robbery at a financial institution is an example of an aggravating factor. One of the few sentencing deductions that defendants are eligible for is a deduction for "acceptance of responsibility."

5. Before *Booker*, a judge could depart from the Guidelines only when the court found that there was "an aggravating or mitigating circumstance of a kind, or to a degree, not adequately taken into consideration by the Sentencing Commission in promulgating the Guidelines." 21A *American Jurisprudence 2d*, §840.

6. The 2005 Sourcebook reports that from October 1, 2004 until January 11, 2005, 70.9% of all federal sentences imposed fell within the guideline range that applied. In the eight and a half months remaining in the court's calendar year after *Booker*, 61.6% of all federal sentences fell within the guideline range. Because the rate of above-the-guideline sentences was under 1%, we know that the increase in non-guideline sentences post-*Booker* consisted largely of below-the-guideline sentences. The difference in below-the-guideline sentences after *Booker* was approximately 10%.

7. A defendant will fall into one of the six criminal history categories on the basis of the number of criminal history points that s/he is assessed as having. These criminal history points are understood as correlating with recidivism. Points are accrued on the basis of features such as previous convictions, conviction of a crime of violence, and misconduct while under judicial supervision. The breakdown of the chart into zones has consequences for the kind of punishment that is imposed: defendants who fall into Zone A are eligible for a non-prison sentence, which could include a period of stay at a Community Corrections Center with work-release privileges, while defendants who fall into Zone B are required to serve their minimum sentence term in prison.

Federal Sentencing Law and Practice lists the following actions that offenders could take that are held as indicating that they have accepted responsibility for their actions:

a. truthfully admitting the conduct comprising the offense(s) of conviction and truthfully admitting or not falsely denying any additional relevant conduct for which the defendant is accountable…;
b. voluntary termination or withdrawal from criminal conduct or associations;
c. voluntary payment of restitution prior to adjudication of guilt;
d. voluntary surrender to authorities promptly after commission of the offense;
e. voluntary assistance to authorities in the recovery of the fruits and instrumentalities of the offense;
f. voluntary resignation from the office or position held during the commission of the offense;
g. post-offense rehabilitative efforts (e.g., counseling or drug treatment); and
h. the timeliness of the defendant's conduct in manifesting the acceptance of responsibility. (2004: 1358)

O'Hear (1997) argues that district and appellate courts tend to have conflicting interpretations of what "acceptance of responsibility" means. O'Hear claims that appellate courts tend to link the acceptance of responsibility benefit to a defendant's remorse, while district courts tend to apply the deduction in response to a defendant's cooperation, i.e., guilty plea. My own observations of district court sentencing hearings lend support to O'Hear's claim regarding the practice of district courts because it was common for defense attorneys to refer to the timeliness of their client's guilty plea (item h) in the process of arguing that their client deserved the a/r deduction.

Defendants can have their total offense level reduced by 2–3 points if the government moves that they have accepted responsibility for their actions. (Three points become available if the total offense level is sixteen or higher (*Federal Sentencing Law and Practice* §3E1.1).) Thus, it is in defendants' interests to be perceived as having accepted responsibility for their actions. If what a defendant says during allocution is inconsistent with this stance, however, the deduction could, in principle, be withdrawn or the judge could decide to sentence the defendant above the minimum that is recommended by the guidelines. Natapoff (2005) discusses a case in which her client made an allocution which called into question the version of events offered by the police. Afterwards the judge accused the defendant of not having accepted responsibility for the incident and imposed a month of imprisonment although the defendant had been eligible for a non-prison sentence.

In my dataset of 52 defendants, 45 (87%) appeared to have received the a/r benefit. Of the seven who did not, four were convicted by a jury. Thus, only three defendants who did not receive the deduction could be understood as having

"lost" this deduction that was potentially available to them because they presumably did plead guilty in a timely manner. At sentencing, special notice was made of the fact that these defendants did not receive the deduction and reasons for withholding the deduction were given: in the three cases, the reasons cited were that the defendants had obstructed justice and/or violated pre-trial release conditions. In sum, although the defendants in this particular dataset largely did receive the a/r deduction, the possibility existed that they could have allocuted in such a way that could lead a judge to withdraw the deduction, or to sentence the defendant above the minimum for which he or she was eligible.

Responsibility for the offense can be communicated in many ways. As O'Hear (1997) and my own observations suggest, entering a guilty plea in a timely manner arguably constitutes the most important way in U.S. district courts. Allocution provides defendants with the opportunity to communicate their acceptance of responsibility verbally. These two means of communicating a/r should not be viewed as equal: without the preceding guilty plea, a defendant's claim to have accepted responsibility at sentencing is likely to have a hollow ring from the sentencing judge's point of view. One way in which speakers can communicate responsibility for an offense in the context of allocution at sentencing is by offering an apology for the offense. Following Goffman (1971), apologies constitute one of several types of "remedies" which function to restore harmony to an interpersonal breach. This study follows Gill in holding that an apology differs from other types of remedies such as excuses and justifications in that it "involves both acceptance of responsibility for the act and an acknowledgment of its wrongfulness" (2000: 12). This understanding of an apology is quite broad. From this perspective, all of the utterances in the dataset in which defendants criticized their actions or offered a personal response to the offense (the B-coded utterances) are understood as being apologetic in nature.

This set of offense-related utterances was divided into eight subcodes, which are presented in Table 1. A token (referred to in column 4) is defined as an utterance consisting of an optional subject and a verb (or verbs if they are conjoined) and its object(s); it is usually demarcated by pauses.

In this paper I describe and analyze the ways in which defendants explicitly position themselves as responsible for their offenses. Thus, I am focusing on the responsibility-oriented (RO) utterances (the lightly shaded, fourth row in Table 1). What makes the RO utterances special is that the defendants appeared to be going out of their way to explicitly demonstrate or claim that they accepted responsibility for their actions. As shown in the table, fifteen defendants produced twenty tokens of utterances that were coded as RO. Twelve defendants produced one RO token; one defendant produced two RO tokens; and two defendants produced three RO tokens.

Table 1. Description of the set of subcodes used to categorize offense-oriented utterances

Subcode name	Description of subcode	No. of defendants who used this code/ percentage of total defendants (n=52)	No. of tokens of code/percentage of all tokens (n=535)
B1 SELF-ASSESSMENT	Defendants describe and/ or criticize their actions and/or the consequences of their actions or they express a wish to undo those actions	27 defendants/52%	66 tokens/12%
B2 EXPLANATION	Defendants offer a reason for why they committed the crime	10 defendants/19%	19 tokens/4%
B3 RESPONSIBILITY	Defendant acknowledges that s/he is responsible for the actions which put her/ him in court	15 defendants/29%	20 tokens/4%
B4 HARM caused to OTHERS	Defendant explicitly acknowledges harmfulness of actions or expresses lack of intention to cause harm to others	9 defendants/17%	18 tokens/3%
B5 SORRY	Defendant uses the word "sorry" to refer to intentional state	29 defendants/56%	43 tokens/8%
B6 FEELING	Defendant uses other feeling-related words (e.g., "ashamed," "regret") to refer to intentional state	7 defendants/13%	12 tokens/2%
B7 APOLOGIZE	Defendant uses primary performative ("apologize") or the corresponding noun ("apology/ies")	17 defendants/33%	28 tokens/5%
B8 FORGIVE	Defendant asks for forgiveness/mercy	5 defendants/10%	7 tokens/1%

6. Formulaic statements of acceptance of responsibility

The RO-coded utterances took two different forms: one was quite formulaic; the other exhibited much more variation in terms of syntactic structure and lexemes. The formulaic option was characterized by the structure: '*I accept/take* (*full*) *responsibility for* NP'; an optional *say*-clause sometimes framed the formula. The '*I accept/take* (*full*) *responsibility for* NP' clause exhibits the syntactic structure of what Austin (1962) identified as performative constructions. The quintessential speech act verb, a performative verb **does** what it says it does; thus, when a speaker

who is authorized to impose sentences says: "It is the judgment of the court that the defendant, ~John Doe, is sentenced to the Bureau of Prisons for a term of 60 months" (in the right context, of course), the defendant is effectively sentenced. As Austin observed, at the performative end of the performative-constative continuum, we find utterances with a subject expressed in the first person singular and a verb in the present indicative mood. Performative utterances are also marked by their ability to co-occur with *hereby*. The '*I accept/take (full) responsibility for* NP' clause is consistent with these criteria and, like other explicit performatives, it also permits the insertion of *hereby*.[8]

Of the fifteen defendants who produced RO-coded utterances, three of them employed a traditional performative structure. (See Appendix A for transcription conventions. To simplify the presentation of data here, pauses of varying lengths are represented by a single (.) surrounded by spaces.)

1. I accept responsibility for what I have done (X9)
2. I take full responsibility for everything, you know, (X10)
3. I accept responsibility for all of my actions, (Z8)

Another three defendants framed their '*accept/take (full) responsibility for* NP' constructions with some kind of preceding clause, two of which were *say* constructions. These utterances have much in common with the traditional performatives described above: they have first person indicative present verb constructions, but the preceding framing material accompanying examples 4–6 makes the insertion of *hereby* unlikely.

4. I just wanna say I take full responsibility for what I've done, (Z3)
5. I'd just like to say that uh, I accept responsibility for what I've done, (Z12)
6. I am totally and solely responsible for my action and accept responsibility for my criminal conduct. (Y6)

In addition to the performative-structured RO-coded utterances, there were two other RO-coded utterances which contained the predicate *take responsibility* but did not share all of the features of the performative structure described above. In example 7, the deleted subject in the *take responsibility* clause refers to a subjunctively-marked *I'd like to* in the previous clause – this clashes with the performative constraint that the verb be in the indicative mood; another feature of divergence is that in both examples 7 and 8, *take* was not the highest verb in the sentence: in these examples verbs of volition occupy the highest syntactic position.

8. To my ears, *hereby* makes a better fit with *accept responsibility* than it does with *take responsibility*.

7. I'd like to apologize for what I've done and take responsibility, (Z6)
8. I have chose to take all responsibility for the money and take all punishment for it . #. and to leave his name out of it for he does not know how I got the money. (X11)

7. Pros and cons of formulaic statements of acceptance of responsibility

The use of the formulaic performative or performative-esque construction is interesting because, while it appears to advance defendants' presumed communicative goals in some ways, it appears to undermine them in others. On one hand, the use of formulaic language is consistent with the formal and ritualistic elements of sentencing hearings and with allocution in particular. Thus, Sadock observes that performative sentences "[a]re commonly used under formal circumstances in Western languages…" (1988: 186). In the 1998 case of *United States v. Myers*, the U.S. Court of Appeals for the Fifth Circuit observed: "… the practice of allowing a defendant to speak before sentencing, which dates back as far as 1689 to the case of Anonymous, 3 Mod. 265, … has symbolic, in addition to functional aspects." The court then cited the case of *United States v. De Alba Pagan* (1994): "As a sister Circuit has observed, 'ancient in law, allocution is both a rite and a right. … Allocution has value in terms of maximizing the perceived equity of the [sentencing] process.'" According to *Myers* and *De Alba Pagan*, as a rite, allocution functions as part of the ceremony of sentencing.

Approaching the ritualistic aspects of sentencing from a different perspective, Garfinkel (1967 [1956]) has highlighted the similarities between sentencing hearings and degradation ceremonies.[9] According to Garfinkel, a "status degradation ceremony" is "[a]ny communicative work between persons, whereby the public identity of an actor is transformed into something looked on as lower in the local scheme of social types." (1967 [1956]: 201) In the process, the character of the person being degraded is recast. His former identity is made to appear as a mere façade and the new, degraded identity is treated as what is real. Garfinkel observes that "the court and its officers have something like a fair monopoly over [these] ceremonies" (1967: 207). If we view sentencing hearings as degradation ceremonies, there are implications for the kinds of defendant statements that exhibit the best fit for this context. For example, defendants who heap blame upon themselves should, in principle, be viewed as embodying a more appropriate stance than those who

9. Thanks are extended to Greg Matoesian for bringing this to my attention.

minimize their blameworthiness. Similarly, defendants who would appear to accept their status as degraded and who do not try to stake out some territory of self that is untainted should be viewed as more closely approaching an ideal defendant stance than those who are seen as trying to claim some moral high ground. From this perspective, the use of formulaic language by the defendant could be understood as indexing acceptance of his/her stigmatized institutional role identity and the shame and remorse that could be understood as accompanying the inhabitance of this role. Interestingly, Judges X, Y, and Z appeared to differ in the degree to which they identified the total person of the defendant as a lawbreaker and wrongdoer and hence, in the degree to which sentencing hearings in their courtrooms functioned as degradation ceremonies. While Judge Y's closing remarks were often sharply critical of defendants, Judge X and Judge Z often used their closing remarks to offer a better picture of life for the defendant in the future. These divergences among only three judges introduce the potential for substantial variation in terms of the kinds of stances that judges might expect from defendants at sentencing, making defendants' task of delivering an effective allocution even more difficult.

In addition, by using a performative structure to communicate their acceptance of responsibility for the offense, defendants appear to be tapping into the force of traditional explicit performatives: by **saying** something is so, speakers make it so. It is fair to assume that defendants want to be viewed as having "accepted responsibility" in order to receive the deduction (or better: not lose the deduction that their attorney and the government have tentatively agreed to). The question arises, though, to what degree *accepting/taking responsibility* for something is an act that is effectively undertaken through speech. The problem is that *taking* or *accepting responsibility* for one's actions can be understood as having both internal and external correlates. Examples of internal correlates include feelings of remorse associated with the awareness of the consequences of one's actions, a commitment to "making things right" to the extent one is able, or a change in one's attitude about the crime, perhaps involving the cessation of denial regarding the events that took place. In contrast, the list of the criteria that the government uses to determine whether an offender is eligible for the a/r deduction consists of external actions. When a defendant explicitly asserts that he has *accepted/taken responsibility* for his actions during allocution at sentencing, it is not clear whether these words are meant to refer to an internal "change of heart" or the external actions enumerated by the government. Unless a defendant elaborates what he/she means by *accept responsibility* in the performative formulaic construction (and defendants rarely did), the phrase is ambiguous. In the context of degradation-ceremonial sentencing hearings, in which defendants are presumed to seek the lowest possible sentence, defendants' use of the ambiguous "accept responsibility" could be viewed as an attempt to use words as a substitute for the actions that they should have taken earlier.

This issue of exactly when the defendant *accepted responsibility* constitutes another problem for the performative and performative-esque formulaic utterances. As Austin observed, performatives with their present tense indicative verbs allow the insertion of *hereby*. Thus part of the power of performatives is that they create their effect at the moment of utterance (assuming uptake by the addressee, of course). However, in order for the a/r deduction to be available to the defendant, the government would have had to recommend it before the start of the sentencing hearing. In fact, as the list of actions used as criteria for awarding the deduction indicate, the taking of responsibility is supposed to begin with the defendant's arrest or shortly thereafter. By foregrounding the here-and-now of the sentencing hearing via the use of a performative construction, an implicit contrast is created between the now (of the hearing) and the then (of the entire period of time between the defendant's arrest and the sentencing hearing). Horn's (1984) work on implicatures is helpful here: because the period of time of the "now" of a speech event is a subset of the period of time that includes events preceding the speech event which are relevant to that event, the choice of a verb that indexes the present will be viewed as implying that the proposition did not hold in the past. Given this potential risk that the use of the present tense presents for the defendant's stance, it is interesting that there aren't any tokens with the form: *I have accepted responsibility for my actions*.[10] Although the numbers are small, the pattern suggests that defendants perceive the plusses of the saying-as-doing performative as outweighing the minuses with regard to chronology.

As noted above, while it appears that the performative and performative-esque construction advances defendants' communicative goals in some respects, in other respects it appears to undermine defendants' performance of a stance of a truly remorseful defendant who is worthy of being given a "break" in sentencing. This problem with the use of the performative stems from the formulaicity of the construction itself. If we start with the premise, following Bakhtin (1981), that words and phrases harken back to other occasions of use, formulaic language should have especially strong connections to those other contexts. Formulaic language produced by defendants will index the myriad events in which other defendants uttered the same or similar words before the sentencing judge. This would appear to have the effect of foregrounding the feature that these speakers have in common – namely, their stigmatized institutional role. In the context of a sentencing hearing, however, it could be argued that it is advantageous for defendants to

10. The closest example is given in example 8 above: Ms. XK's, *I have chose to take all responsibility for the money and take all punishment for it…* (X11), where she employs the present perfect, but the highest verb in the sentence is *choose* instead of *take*.

highlight their individual identity over their institutional identity where possible. In order to receive a sentence that is different from that of the majority of defendants (which, currently, averages to about five years), defendants would presumably want to present themselves in court in as individualized a manner as possible.[11]

There is another potential problem linked to defendants' use of formulaic performatives in this context. Austin's identification of performative utterances as quintessential speech acts provides a link between forms and speaker intentions. In the context of a sentencing hearing, which is structured around the event of imposing a sentence (typically of imprisonment in a Federal Correctional Institution), the defendant's presumed communicative goals will have a default association with that sentence: that is, defendants will be presumed to desire the minimum sentence that is available. As a result, defendants who use Austinian performative language at sentencing are likely to have the goals of their speech acts understood with respect to the upcoming sentence – specifically as attempting to reduce that sentence. The inter-textual links between the name of the "acceptance of responsibility" sentencing deduction and defendants' use of *responsibility* and especially *accept responsibility* function to make this association even stronger. The self-interest that a defendant has in this regard could diminish his or her capacity to be perceived as sincerely remorseful (cf. Gruber in press).

8. Less formulaic statements of acceptance of responsibility

Using performative and performative-esque constructions constituted one way in which defendants expressed explicit acceptance of responsibility for their actions. As noted earlier, there was another way in which defendants accomplished this same goal. This way was marked by much more varied kinds of expressions:

1. I know that life is an open book and on each page is written the choices I have made. (X1)
2. My actions are the sole reason. I am **here** today. No one put me here but me. (X1)
3. It is by all means my own wrong mistakes and choices that have gotten me in this situation. (X9)

11. The average sentence for the fifty-two defendants in the dataset was 79.2 months and the median sentence length was 58.8 months. According to U.S. government statistics, between October 1, 2004 and September 30, 2005, the average length of a federal sentence nationally was 61.2 months. This number excludes defendants sentenced to life, death, and other anomalous circumstances. (See: http://www.uscourts.gov/judbus2005/appendices/d5.pdf.)

4. I –. you know, I know I did this to myself, (X16)
5. Still, #####; I still went along with it. (Z9)
6. I mean that's no excuse for what I've done. (X5)

9. Pros and cons of less formulaic statements of acceptance of reponsibility

When we compare the ways in which defendants refer to their criminal actions across the formulaic and less formulaic groups, the lexicon they employ is very similar; that is, defendants frequently refer to their crimes by means of phrases such as *my actions* and *what I've done*. In fact, the majority of the allocutions in the dataset give no information as to the specific crime for which the defendant was being sentenced (Gruber 2007). One way in which the two groups of RO-coded utterances differ, however, is the grammatical slot in which the reference to the crime appears. In the more formulaic set of RO-coded utterances, defendants' use of *take responsibility* or *accept responsibility* requires a *for*-prepositional phrase. The reference to the crime, then, is accomplished by means of an NP in a prepositional phrase. In the more creative RO-coded utterances, defendants make use of a variety of grammatical positions to refer to their crimes: as the subject: *My actions are the sole reason I am here today*; as the predicate: *I did this to myself*; as well as in prepositional phrases: *that's no excuse for what I've done*. It is a feature of English grammar that agentive NP's occur in subject and object positions much more often than they do in the NP slot of a prepositional phrase. Thus, referring to the crime in more grammatically agentive ways appears to function iconically to index the defendant as performing a stance of taking responsibility rather than simply producing a verbal claim of doing so. In this way, defendants avoid the problem of the ambiguity between thoughts and deeds created by the use of the formulaic *accept/take responsibility* constructions discussed above.

Support for the idea that more creative, less formulaic utterances might better serve defendants' communicative goals at sentencing finds echoes in Johnstone's (1991) work on public opinion surveys, Sugimoto's (1998) work on apologies and Shoaps' (2002) work on Assembly of God Church prayer. Johnstone argued that telephone interviewers who added unscripted elements to their interchanges were better able to "point up their identities as individuals rather than merely fillers of the interviewer role" (1991: 557). Such a strategy actually helped them to be more successful in persuading interviewees to participate in and complete the survey. Sugimoto (1998) found that, in contrast to Japanese etiquette books, American etiquette books exhorted readers to use more creative apologies because they sounded more sincere than formulaic ones. And Shoaps (2002) found that Pentecostal

church communities tend to view creatively elaborated prayers as reflecting more earnest prayer than formulaic ones. This body of work suggests that there is a palpable link in American culture between the perception of more creative language and speaker sincerity in the performance of expressive speech acts.

Just as there were pros and cons associated with the use of formulaic language, the use of more creative language is not without its potential drawbacks. Defendants who use more creative constructions are unable to tap into the positive features associated with the formulaic constructions: by highlighting their individuality, they simultaneously downplay the degree to which they index their occupancy of a stigmatized institutional role identity. In courtrooms in which sentencing hearings more closely approximate degradation ceremonies, such a communicative strategy may cause them to be viewed as being in denial of the institutional role that they now inhabit. From this perspective, using language with strong inter-textual links to the impending sentence could be understood as indexing defendants' awareness of the reality of their situation. This closer look at the challenges faced by defendants in communicating acceptance of responsibility during allocution at sentencing suggests that no single strategy involving formulaic or less formulaic language will allow defendants to achieve all of their communicative goals.

10. Conclusions and implications

An examination of defendants' statements of acceptance of responsibility for their actions reveals that both formulaic and non-formulaic options advance and undermine defendants' presumed communicative goals at sentencing. As Gruber (in press) observes, the double binds which speaking on their own behalf presents to defendants sharply contrasts with the language ideologies surrounding allocution. In this paper I propose that in sentencing hearings – especially those in which the total person of the defendant is identified as a wrongdoer as in degradation ceremonies – the use of formulaic language could index a defendant's acceptance of his/her stigmatized institutional role identity, which could have implications for positively-viewed traits that are understood as accompanying this role, such as remorse. Further, these particular formulaic examples exhibited clear ties to Austinian performatives, which add connotations of agentivity to defendants' statements. As was noted, however, defendants who use formulas to communicate their acceptance of responsibility risk foregrounding their similarities to other defendants who have produced the same formulas. In this way, they diminish their capacity to present themselves in an individualized manner. In addition, the use of formulas that contain the lexical items "accept" and/or "responsibility" which appear in the name of a potentially-applicable sentence deduction functions to

foreground a defendant's self-interest in the upcoming sentence. This presumed self-interest undermines the sincerity of a defendant's statement of responsibility.

While the use of more creative language to "accept responsibility" avoids the pitfalls associated with formulaic language and appears to provide other benefits as well, it is also not without potential risks. From the positive side, an examination of more creative RO "acceptance of responsibility" allocutory strategies suggests that references to criminal actions that occur in more grammatically-agentive sentential slots could serve to index the speaker as having more fully "accepted responsibility" for his or her actions. Further, a growing body of work suggests that there is a cultural link between creativity and perceived sincerity in the performance of expressive speech acts. However, we also noted that speaking in more creative ways could index the defendant as unwilling or unable to inhabit the institutional role identity of criminal defendant. Although an intuitive calculation of the pros and cons associated with the choices available to defendants for expressing acceptance of responsibility appears to perhaps favor the non-formulaic constructions over the formulaic ones, a closer look reveals that there are no formulas for success in this context.

The meanings that are associated with the use of (relatively) formulaic vs. non-formulaic language in the context of allocution contrast strikingly with the meanings associated with formulaic vs. non-formulaic language in the literature on implicatures and politeness (cf. Grice 1957; Brown & Levinson 1987 [1978]). Thus, for example, Wray (1999) discusses the use of formulaic expressions such as *excuse me* in the context of trying to maneuver through a noisy, crowded bar. She observes that the formulaic expressions are easily recognized as non-confrontational requests. "In contrast, a less formulaic utterance such as *I'm walking behind you* must be heard more accurately because it is unpredictable and requires more decoding." Thus, relative to "*excuse me*," "*I'm walking behind you*" is heard as a more intrusive request. (1999: 216) This example is consistent with Grice's Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (1991 [1957]: 307), in particular the second maxim of Quantity: "Do not make your contribution more informative than is required." (1991 [1957]: 308).

Making an apologetic statement during allocution at sentencing is different from trying to maneuver through a crowded bar or engaging in one of Grice's straightforward examples, such as asking for four screws in the process of repairing one's car – and not only because it takes place in a courtroom. If we assume, following Sadock (1994) among others, that language use ordinarily involves three separate communicative aspects – the informational, effective (or social), and affective aspects – and that speech act types "differ precisely in which of the three basic components is the principle component" (1994: 398), Wray's bar-hopper and

Grice's handyman are engaging in communicative tasks that privilege the effective or social aspect. Making an apologetic statement on one's own behalf privileges a different communicative aspect: the affective component. Another example of a speech act that privileges the affective component is sympathizing with someone. Thus, Austin's (1962) fundamental idea that different kinds of communicative tasks involve different kinds of actions via language has implications for the ways in which formulaic language vs. non-formulaic language will function interactively. If we limit the scope of our investigation of formulaic language to a single type of communicative exchange, we will miss much of the richness of the ways in which the continuum of choices ranging from formulaic to non-formulaic language have different meanings in different contexts.

The inter-textuality of formulaic language makes it an extremely rich site for analysis. This study highlights the importance of considering the ways in which the institutional role identity of the user of formulaic vs. non-formulaic language both create and limit opportunities for meaning-making in a particular context. While the institutional role identity of criminal defendant was shown to impact the meanings that were associated with the use of formulaic or non-formulaic language during allocution at sentencing, the findings presented here have implications beyond the courtroom. Just as congruences of meanings of formulaic vs. non-formulaic language are seen to transfer across different contexts in which the social or effective aspect of communication is privileged, we expect that patterns of meanings of formulaic vs. non-formulaic language will be evident across contexts that privilege affective meaning as well.

Appendix A

Transcription symbols (largely following Du Bois 2006)

~XA	name or initial change to preserve anonymity (tilde)
#	unintelligible; one per syllable
(1.2)	pause duration in seconds and tenths of seconds
..	< 0.2 seconds; silence, break in phonation
word	bold-face indicates emphasis via loudness or contrastive pitch

Sample Allocution

Defendant: ~XP; Age (at time of offense): 26; male; African-American; level of education: 11 years. Charge: Controlled substance – sell, distribute, or dispense (count 1); Violent crime/drugs/machine gun (count 2); Receive stolen firearms (count 3). Pleaded guilty to counts 1 and 2. Total Offense Level 10, Criminal History Level I (Received 2 level deduction for a/r). Guideline Range for Count 1: 6–12 months; Count 2 has statutory minimum of 60 months. Sentenced to a total of 66 months.

In court for defendant: 13 people, one of whom is ~Mr. P's pastor

X16. ~Judge X; ~Mr. P, would you like to say anything on your own behalf before I sentence you?

1. ~Mr. P; A2 (.6) Uh, yes, Ma'am, I would. (.7)
2. E1 I'd just like to um, (.8) uh, (.6) first thank my family and friends and (.6) um (.8)
3. people for supporting me, you know, through this. (.8)
4. D3 And .. it's been a real hard time for me and them also,
5. B3 I – (1.2) you know, I know I did this to myself,
6. B1 I made mistakes .. and (.6)
7. C1 I've definitely learned from it, you know. (1.2)
8. C2 I guarantee myself that I would – (1.3) next time anyone hears about me or
9. anything like that, it's .. gonna be in a positive manner, you know, in a
10. positive direction, so (.6)
11. H2 I #don't have a lot to say.
12. G2 Thank you.

~Judge X; (3.1) (THROAT) Thank you.

Appendix B: Federal Sentencing Table (2006 Federal Sentencing Guidelines Manual)

SENTENCING TABLE
(in months of imprisonment)
Criminal History Category (Criminal History Points)

	Offense Level	I (0 or 1)	II (2 or 3)	III (4,5,6)	IV (7,8,9)	V (10,11,12)	VI (13 or more)
	1	0–6	0–6	0–6	0–6	0–6	0–6
	2	0–6	0–6	0–6	0–6	0–6	1–7
	3	0–6	0–6	0–6	0–6	2–8	3-9
	4	0–6	0–6	0–6	2–8	4–10	6–12
Zone A	5	0–6	0–6	1–7	4–10	6–12	9–15
	6	0–6	1–7	2–8	6–12	9–15	12–18
	7	0–6	2–8	4–10	8–14	12–18	15–21
	8	0–6	4–10	6–12	10–16	15–21	18–24
	9	4–10	6–12	8–14	12–18	18–24	21–27
Zone B							
	10	6–12	8–14	10–16	15–21	21–27	24–30
Zone C	11	8–14	10–16	12–18	18–24	24–30	27–33
	12	10–16	12–18	15–21	21–27	27–33	30–37
	13	12–18	15–21	18–24	24–30	30–37	33–41
	14	15–21	18–24	21–27	27–33	33–41	37–46
	15	18–24	21–27	24–30	30–37	37–46	41–51
	16	21–27	24–30	27–33	33–41	41–51	46–57
	17	24–30	27–33	30–37	37–46	46–57	51–63
	18	27–33	30–37	33–41	41–51	51–63	57–71
	19	30–37	33–41	37–46	46–57	57–71	63–78
	20	33–41	37–46	41–51	51–63	63–78	70–87
	21	37–46	41–51	46–57	57–71	70–87	77–96

(*Continued*)

Appendix B: (Continued)

	Offense Level	I (0 or 1)	II (2 or 3)	III (4,5,6)	IV (7,8,9)	V (10,11,12)	VI (13 or more)
	22	41–51	46–67	51–63	63–78	77–96	84–105
	23	46–67	51–63	57–71	70–87	84–105	92–115
	24	51–63	57–71	63–78	77–96	92–115	100–125
	25	57–71	63–78	70–87	84–105	100–125	110–137
	26	63–78	70–87	78–97	92–115	110–137	120–150
	27	70–87	78–97	87–108	100–125	120–150	130–162
Zone D							
	28	78–97	87–108	97–121	110–137	130–162	140–175
	29	87–108	97–121	108–135	121–151	140–175	151–188
	30	97–121	108–135	121–151	153–168	151–188	168–210
	31	108–135	121–151	135–168	151–188	168–210	188–235
	32	121–151	135–168	151–188	168–210	188–235	210–262
	33	153–168	151–188	168–210	188–235	210–262	235–293
	34	151–188	168–210	188–235	210–262	235–293	262–327
	35	168–210	188–235	210–262	235–293	262–327	292–365
	36	188–235	210–262	235–293	262–327	292–365	324–405
	37	210–262	235–293	262–327	292–365	324–405	360–life
	38	235–293	262–327	292–365	324–405	360–life	360–life
	39	262–327	292–365	324–405	360–life	360–life	360–life
	40	292–365	324–405	360–life	360–life	360–life	360–life
	41	324–405	360–life	360–life	360–life	360–life	360–life
	42	360–life	360–life	360–life	360–life	360–life	360–life
	43	life	life	life	life	life	life

References

2006 *Federal Sentencing Guidelines Manual*. U.S. Sentencing Commission.

2005 *Sourcebook of Federal Sentencing Statistics*. U.S. Sentencing Commission.

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

American jurisprudence, 2nd Edn. 1998. Rochester NY: Lawyers Cooperative Publishing Company.

Austin, John L. 1962. *How to do things with words*. Cambridge MA: Harvard University Press.

Bach, Kent & Robert P. Harnish. 1979. *Linguistic communication and speech acts*. Cambridge MA: The MIT Press.

Bakhtin, Mikhail. 1981. The dialogic imagination. M. Holquist (Ed.), Austin TX: University of Texas Press.

Brown, Penelope & Stephen C. Levinson. 1987 [1978]. *Politeness. Some universals in language use*. Cambridge: CUP.

Coulmas, Florian (Ed.), 1981. *Conversational routine*. The Hague: Mouton.

Du Bois, John W. 2006. Representing discourse. (ms.) Available at http://www.linguistics/ucsb.edu/projects/transcription/representing.

Federal criminal code and rules. 2004 Edn. St. Paul MN: West Group.

Garfinkel, Harold. 1967 [1956]. Conditions of successful degradation ceremonies. *Symbolic interaction. A reader in social psychology,* 2nd Edn. J.G. Manis & B. Meltzer (Eds), 201–208. Boston MA: Allyn and Bacon.

Gill, Kathleen. 2000. The moral functions of an apology. *The Philosophical Forum* XXXI (1): 11–27.

Goffman, Erving. 1971. *Relations in public. Microstudies of the public order*. New York NY: Basic Books.

Greenhouse, Linda. 2007. Justices to revisit thorny issue of sentencing guidelines in first cases after recess. *New York Times*, February 20, 2007.

Grice H. Paul. 1957. Meaning. *The Philosophical Review* 66(3): 377–388.

Grice, H. Paul 1991 [1968]. Logic and conversation. In *Pragmatics: A reader*, S. Davis (Ed.), 305–315. Oxford: OUP.

Gruber, M. Catherine. 2007. A linguistic and ethnographic analysis of apology narratives performed in the context of federal sentencing hearings. Ph.D. dissertation, University of Chicago.

Gruber, M. Catherine. 2008. Contextual constraints on defendants' apologies at sentencing. *Studies in Law, Politics, and Society* 45: 47–74.

Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In *Georgetown University Roundtable on Languages and Linguistics*, D. Schiffrin (Ed.), 11–43. Washington DC: Georgetown University Press.

Hutchinson, Thomas N., David Yellen, Peter Hoffman & Deborah Young (Eds) 2004. *Federal sentencing law and practice*. St. Paul MN: West Group.

Johnstone, Barbara. 1991. Individual style in an American public opinion survey: Personal performance and the ideology of referentiality. *Language in Society* 20: 557–576.

Natapoff, Alexandra. 2005. Speechless: The silencing of criminal defendants. *New York University Law Review* 80(5): 1449–1504.

O'Hear, Michael. 1997. Remorse, cooperation, and 'acceptance of responsibility': The structure, implementation, and reform of Section 3E1.1 of the Federal Sentencing Guidelines. *Northwestern University Law Review* 91: 1507–1573.

Pawley, Andrew. 2007. Developments in the study of formulaic language since 1970: A personal view. In *Phraseology and culture in English*, P. Skandera (Ed.), 3–45. Berlin: Mouton der Gruyter.

Sadock, Jerrold. 1988. Speech act distinctions in grammar. In *Linguistics: The Cambridge survey,* Vol. II: *Linguistic theory: Extensions and implications*, F.J. Newmeyer (Ed.), 183–97. Cambridge: CUP.

Sadock, Jerrold. 1994. Toward a grammatically realistic typology of speech acts. In *Foundations of speech act theory*, S.L. Tsohatzidis (Ed.), 393–406. New York NY: Routledge.

Shoaps, Robin A. 2002. 'Pray earnestly': The textual construction of personal involvement in Pentecostal prayer and song. *Journal of Linguistic Anthropology* 12(1): 34–71.

Sugimoto, Naomi. 1998. Norms of apology depicted in U.S. American and Japanese literature on manners and etiquette. *International Journal of Intercultural Relations* 22(3): 251–276.

Tavuchis, Nicholas. 1991. *Mea culpa. A sociology of apology and reconciliation*. Stanford CA: Stanford University Press.

Wray, Alison. 1999. Formulaic language in learners and native speakers. *Language Teaching* 32: 213–231.

Wright, Charles Alan, Nancy J. King & Susan R. Klein (Eds), 2004. *Federal practice and procedure.* Thomson/West.

Cases Cited

Green v. United States, 365 U.S. 301 (1961)
McGautha v. California 402 U.S. 183 (1971)
United States v. Booker, 125 U.S. 738 (2005)
United States v. De Alba Pagan, 33 F.3d 125 (1994)
United States v. Myers, 150 F.3d 459 (1998)

Decorative symmetry in ritual (and everyday) language

John Haiman & Noeurng Ourn
Macalester College

Abstract

Symmetrical compounds like *last and final, jibberjabber* and *pell-mell* exist in English, but are much more richly attested, especially in the formulaic registers, of Southeast Asian languages like Khmer. The consensus of opinion is that the repetition serves a decorative rather than a cognitive referential function. There are four kinds of evidence that in Khmer the nonsense (= *jibber*) portions of alliterative or rhyming compounds "have no etymology": rather than deriving from meaningful roots via erosion, they are made up to satisfy a "Drang nach Parallelismus" for purely esthetic imperatives. The same drive may be attested in agreement, aggressive reduplication, structural priming, and even baby talk in other languages.

1. Introduction: Galumphing, non-referential bulking and decorative symmetry in Khmer

In everyday life, ritual or public behavior is frequently marked by ornamentation. In fact, the degree of ritualization of almost any pragmatic performance can often be directly read from the number of non-functional bells and whistles that adorn it. (Compare marching with walking, high fashion with casual clothing, a formal dinner with eating at home alone, and so on.) Ordinary language is exactly comparable to other kinds of ritualizable behavior. A speech is not like ordinary conversation. Often this contrast is quantifiable. For example, it is a solidly established finding that politeness and high register in every language contrast with informal and/or casual speech in the amount of non-referential "galumphing" (Miller 1973) that encumbers the referential message. It is no accident that four letter words in English are four letter words, while the corresponding euphemisms are longer. Geertz's 1955 classic presentation of levels of speech in Javanese is still one of the richest attestations of this familiar phenomenon: the non-speaker of Javanese can in effect deduce the six different respect levels from lowest to highest for a single utterance ("Are you going to east rice and cassava now?") by simply counting the total number of syllables in each version.

One of the most popular ways of providing "non-referential bulking" in stylized ritual language is by substituting PAIRS of parallel expressions for SINGLE expressions (Boas 1925; Queneau 1947; Jakobson 1966; Fox 1988; Bright 1990; Wälchli 2005). The pairs may be synonyms or near synonyms, as in *rules and regulations, last and final, let or hindrance, complete and utter, in this day and age* and so forth. Cambodian, a Mon-Khmer language, is particularly rich in a variety of examples of this type (as are many other languages of the Southeast Asian linguistic alliance), and this essay will be largely devoted to a description and analysis of the devices that are dragooned into the expression of what we are calling decorative or non-referential symmetry.

In contrast with reduplication, emphatic repetition, and other formally related devices, the repetition in cases of decorative symmetry seems to serve little or no referential function. The ornamentation is not an icon of plurality, intensity, repetition, habituality, or any of the other things that can and often are signaled by formal repetition (cf. Wierzbicka 1987; Farghal 1992 for tautologies like *East is East*; Sapir 1921: 76–7 and Moravcsik 1978 for reduplication in general; Heine & Reh 1984: 46–7 for reduplication in African languages, Svantesson 1983: 115–9, Khin 2002: 119–20 for reduplication in Mon-Khmer languages including Khmer). In fact, it is kept rigidly separate from referential repetition in the following remarkable way: verbatim and literal repetition of a form is *reserved* in Khmer for

the expression of plurality, repetition, intensity, or tautology, and apparently can never perform a purely decorative function. For example *good good pig* in English may mean "very good pig", *water, water* may mean "lots of water", *tried and tried* may mean "kept trying" and so on. In Khmer *cruuk l'aaw l'aaw* ('pig good good') means "good pigs", *muh nih muh* (literally 'mosquito this mosquito') means "lots of mosquitoes", *awh kaaw awh teuv* (literally 'exhaust and exhaust go' means "gone is gone", and *awngkuj awngkuj* ('sit sit') means "keep trying to sit down". But, as these examples indicate, the literal combination/X X/always iconically signals an intensification or repetition or pluralization of "X", never simply "X by itself". In contradistinction, the purely decorative near repetition (almost) always does signal "X by itself" and nothing more. The value added by the repetition is thus purely esthetic, and not referential.

In some languages, non-referential symmetry of this sort is barely attested, while in others, it is an inescapable feature of good style. Khmer, like many other languages of Southeast Asia, has exaggerated this tendency towards parallelism to a remarkable degree. Consider the following largely formulaic text, a New Year's Greeting:

Soom aoj baan seckdej sok {cawmrong cawmraeun}
Wish that get matter peace [Twin form] plenty

{Prawkaawp daoj} sokha'phiap baw'ri'boo {prah craak}
with through good health/prosperity sufficient escape flee

{cumngww chww tkat} {crah srawlah} awmpii {tok soak}
illness illness clear clear from suffer weep

{mae mav} {kdav krawhaaj} haeuj soom {baan tautual} neuv phiap
frustration hot burning hot and wish get receive obj. aspect

{sngawp rumnoap} {sav maawng} soom aoj {prawkaawp neung}
calm pacify distress sorrow wish that with with

phiap {sokdom rau'mania} sawnta' phiap ni'raun taaw teuv
aspect bliss happiness peace eternal continue go

Freely translated: '(I) wish (you) peace, prosperity, tranquility, freedom from suffering, serenity, the surcease of sorrow, and (I) wish (you this) with eternally continuing bliss and happiness.'

This is a written New Year's salutation, a presumably typical example of a genre which in its portentousness is typical of ritual speech. Non-referentially parallel structures are those which are bracketed in {..}. As Geertz famously said in his discussion of "linguistic etiquette" in Javanese, "there is a peculiar obsession at work here" – something "peculiar" at least, to speakers of English. It is, incidentally, one that is recognized by speakers of Khmer themselves. A folk legend tells of how the Buddha blessed and fed the *trawkuat,* a monitor lizard who

was present at a sermon that he gave in *Kook Tlaawk,* the land that would become Cambodia. When his disciples asked why the lizard was so honored, he answered that the people who would live here, like the lizard, would have forked, but innocent tongues, forever saying things twice.

One of the ways in which Cambodians can say things twice, is to couple a word with a near synonym. English, French, and German do just barely exhibit this tendency, (witness *last and final* etc.) but a number of other European languages like Georgian and Mordvin are closer to Khmer in their exuberance of such synonym compounds (cf. Wälchli 2005). A very (very!) fragmentary exemplification is provided in the New Year's greeting above. In Khmer, the range of sources for such synonyms includes foreign languages like Pali or even French:

(1) a. *lumneuv* residence (Khmer) *thaan* residence (Pali) "residence"

b. *kdav* warm (Khmer) *un* warm (Pali) "warm, hot"

c. *caekcee* chercher (French) *raauk* seek (Khmer) "seek"

or a borrowing from Pali may be conjoined with its doublet, the same word borrowed from Sanskrit:

(2) a. *peel* time (Sanskrit) *weelia* time (Pali) "time"

b. *teevaudaa* god (Sanskrit) *teep* god (Pali) "angel, good spirit"

c. *rak* guard (Pali) *reaksaa* guard, (Sanskrit) "guard" take care of

While English allows coupling of synonymous nouns, verbs, and adjectives, Khmer allows the coupling of conjunctions and prepositions as well: for example *baeu* and *prawseun* both mean 'if' and may be conjoined in either order – without necessarily meaning (say) 'if and only if' or anything other than simply 'if'. The New Year's text includes {*prawkaawp daoj*} as the double representation of the preposition "with".

A major (indeed perhaps the only) significant limitation on synonym compounds of this sort is that synonyms taken from different registers (e.g., Everyday, High, Very High) may not be coupled. For example *chang* "eat" (to or about a monk) may never be coupled with *sii* "eat" (to or about an animal).

2. The primacy of phonetic motivation for decorative servant words

Another non-referential symmetrical device is to couple a word with a "twin form" (Marchand 1960) or a *bo'ri'vaa sap* 'servant word' (Ourn & Haiman 2000), a morpheme that seems to have no independent existence. A single attestation is encountered in the New Year's text above, the combination {*cawmraeun cawmrong*} – whose first constituent means "plenty", and whose second constituent seems to mean nothing at all. Not only English, but many other languages have such parallel, somehow "expressive" words (which are sometimes classed with ideophones): *jibber jabber, flimflam* or *helter skelter* (cf. Pott 1962; Paul 1880: 181, for the phenomenon in general, Marchand 1960, chapter 9 for twin forms in English, Nacaskul 1976 for a survey of four unrelated Southeast Asian languages, Nguyen 1965 for Vietnamese, Weidert 1973 for Khasi, another Mon-Khmer language, Gregerson 1984 for Rengao, Vongvipanond 1992 for Thai, Ratliff 1992: 136–45 for Hmong, Stanford 2007, to appear for Sui). Working within a tradition of referential motivation, a number of researchers in SE Asian languages have grappled with the raison d'être for compounds of this sort.

First of all, the phenomenon is recognized by Khmer grammarians, who note explicitly that the decorative *bo'ri'vaa sap* ('servant word') is meaningless (see, for example Chaaun Chiang 2002: "the servant word alone has no meaning by itself".)

Rischel (1995: 93–4) notes the existence of "doubling" in expressive words in Minor Mlabri, a Mon-Khmer language of Northern Thailand, but adds that

> It is conspicuous that the doubling phenomenon is not restricted to specifically expressive words. It even occurs on nouns referring to ordinary physical objects … [including] body parts such as *klkiil* "knee", *mujmuj* "hair" and certain terms for age and sex categories such as *burbur* "young man"… There may be a very interesting explanation, but what is it?

Svantesson (1983: 124–5) recorded nearly 400 examples of what he called "reduplicatives" in Kammu, another Mon-Khmer language, and claimed that they all had "an intensifying function" but it is difficult to see what part of the meaning of words like "one-eyed", "mountain side", or "gift" is "intensified" in forms like *'yee'yang (<'yee), k'aar k'ir (<k'aar),* or *pntrap pntreeng (<pntrap).*

Weidert (1973: 141) acknowledged that in (Mon-Khmer) Khasi the meanings of "redundant compounds" like *mdan mdia* 'meadow', *tlawt tlar* 'weak', and *prhut prham* 'wind' are except for very rare cases exactly the same as the roots from which they are formed (*mdan, tlawt, prhut*).

In his grammar of Thai, Smyth (2002: 97) observed that "it is sometimes difficult to distinguish any real difference in meaning between a single and a reduplicated form; in cases where the reduplicated form is preferred, it seems to be because it creates a rhythm that is more pleasing to the ear."

Roffe (1975: 285) claimed that in the closely related Lao language [binomial expressions of this sort] :

> are to the language what spice is to food, what polishing or cutting is to a gem. Without them, the speaker or writer will make himself understood, but prove to be rather dull or pedestrian.

Nacaskul (1976: 874–6) specifically claimed that base forms... in Thai can be inflated without altering their lexical meaning… This is a simple esthetic datum that it behooves us to analyze and which does not have to justified…casual speech.. purged of elaborations, although unambiguous and grammatically correct, sounds unpleasantly harsh and alien – in fact, we would claim, much like "translationese": that is, wellformed, but simply not idiomatically natural.

Stanford (2007, to appear) in two studies of such words in Sui (another Thai-Kadai language, spoken in Southwestern China) calls the echo portion of each word an "intensifier": *khing* 'brown', *khing ting* 'very brown'. This begs the same question as Svantesson's characterization of such intensifiers in Kammu. If *cu* means 'green' for example, how does one justify that *cu cing* means 'very green'? Is it not the next thing to saying that it means practically nothing at all?

In Hmong, double words are expressive ideophones like English *splat*: that is, they can neither be translated into propositional language (Ratliff 218, fn) nor negated or questioned (ibid.. 138), and they are frequently introduced by *nrov* 'sound, go' or follow a verb like "fall", "move" or "cry", which they then can be thought to "modify" (ibid.139). The meaning of the alliterative echo portion (which always precedes rather than follows the root morpheme, ibid. 142) is problematic. Ratliff (ibid. 137) cites a Chinese scholar (Ts'ao 1961; translated in Purnell 1972), whose reactions echo those of other observers of this phenomenon:

> These syllables are sometimes incomprehensible to the Chinese comrades who study Miao. Even when they do understand them, it is very difficult to translate the Miao meaning into Chinese… Without such syllables the language would be much less colorful.

In Ourn & Haiman 2000, appendix 5, we provided a provisional catalog of several hundred such forms in Cambodian, and said nearly nothing about their function(s).

Clearly, words like *jibber jabber* often have some iconic and expressive function not shared by the single words from which they are sometimes seen to derive.

Wälchli (2005: 126) draws attention to the idea of "disordered iterativity" in English words like *helter-skelter, hurly burly, hodgepodge, pell-mell, higgledy piggledy, mishmash*, paralleled by the same idea in German expressive compounds like *Mischmasch, Krimskrams* 'assorted junk', *Wirrwarr* 'confusion, jumble', *Kuddel Muddel* ('idem'), *Schnickschnack* 'balderdash'. Compare Mikone (2001: 230) who finds a common meaning of "sloppy"-ness in a number of symmetrical compounds in Estonian: *liga loga* 'confused, sloppy, bad', *pira para* 'scattered, sloppy, careless'. A case could be made for similar expressivity in a number of Khmer cases like *rauhah rauhuan* 'quick' (from *rauhah*). But in the vast majority of such decorative compounds, we find ourselves facing the same issue that confronted researchers like Chiang, Rischel, Roffe, Smyth, Svantesson, Weidert, Ratliff, Stanford, and Nacaskul. The extra material, while colorful, just doesn't seem to mean anything.

It is natural to suppose, via the "junkyard" or "compost" metaphor with which we are familiar from traditional studies of grammaticalization, that presently meaningless decorative reenforcements of referential words are the decayed remnants of originally rhyming or alliterating near-synonyms which have tended over time to

a. lose their autonomy, much like the presently meaningless reenforcers *kith* and *Kegel* in expressions like English *kith and kin* or German *Kind und Kegel*. (cf. Hock & Joseph 1996: 169; Ourn & Haiman 2000: 483).
b. grow phonetically more similar to the words they are conjoined to, the more so as they lose their meanings, perhaps along the lines of *caboodle* (< *boodle)* in expressions like *kit, cat, and caboodle* (cf. Wälchli 2005: 148).

A much more radical hypothesis, however, is that these meaningless forms are or were made up on the spot and out of whole cloth for purely formal reasons, and therefore have no etymologies at all. (Thus Maspero 1915: 226; Gorgoniev 1966: 73; Chiang 2002: 10). Purely negative evidence for the "whole cloth" hypothesis is the (non-) finding that etymologies for most twin forms are usually very hard to find, even in relatively overstudied Indo-European languages like French, English, and Spanish. Malkiel (1970 : 353) is able to offer only a couple, for *pêle-mêle*, and *bric-à-brac*.

There are, as we might expect, a number of Khmer forms of indeterminate status. Not only are there a fair number of existing synonym pairs which happen to alliterate, just as most servant words do, like *cah cauria* "old decrepit" (cf. Ourn & Haiman 2000, appendix 2). So, too, sources and authorities may also disagree on whether an ambiguous form is a "servant word" or a near synonym: for example the first element of *pkoap pkuan* "satisfy" may be meaningless (hence a servant word) as suggested in the dictionary of Chuon Naath, first published in 1938; or it may be a possible causative *p-koap* of the verb *koap* 'be pleased', as seems likely to Noeurng. The existence of "grey areas" is however

equally compatible with either the junkyard theory (according to which these forms have not yet fully disintegrated) or the whole cloth theory (which maintains that these forms may have not yet fully emerged as independent words).

The whole cloth theory asserts that servant words are created exclusively for their formal phonetic properties. There are at least four sorts of evidence which argue for the correctness, or at least the high plausibility, of this theory.

2.1 Compounding by conscription

First, some functionally meaningless servant words are meaningful elsewhere in the language. But in the decorative compounds where they appear, the meaning that they have elsewhere is totally irrelevant, suggesting that they were conscripted for their decorative function for purely phonetic reasons.

Consider the following compounds, whose referentially relevant member is glossed in capitals below

lveung	*lviaj*	"vast"
VAST	slow	
bawnlae	*bawngkaa*	"vegetables"
VEGETABLES	protect	
bawnlae	*bawnlawm*	"confuse, distract"
Vegetables	CONFUSE	
kawmlang	*kawmhaeng*	"force, energy"
FORCE	yell	
rauliing	*rauloong*	"clear, transparent"
Empty	CLEAR	
sawmdej	*samdav*	"speech"
SPEECH	towards	
prakaawt	*prawcia*	"exact"
EXACT	person	
tnak	*tnaawm*	"handle carefully"
level	HANDLE	
psaawp	*psaaj*	"disseminate, propagate, spread"
perception	spread	

In these examples, the junior partner in each pair might as well be meaningless, and was initially perceived as meaningless by Noeurng.

2.2 Compounding by prosthesis

In other pairs, both elements are meaningful, but one of them has been tricked out with a semantically empty prefix or infix to make it alliterate with the senior partner. We have seen this in the English example *kit, cat, and ca-boodle.* The meaningless affix is represented in capitals in the examples below:

PRAW-hak	*prawhael*	"like, about, approximately"
like	approximately	
mhoop	*M-haar*	"food"
food	food	
d-AWM-kom	*dawmkaeung*	"carry up"
gather	carry up	
srawngoot	*s-RAW-ngat*	"melancholy"
sad	quiet	
DAWNG-hoo	*dawnghae*	"parade, procession"
flow	parade	

2.3 Compounding via the "Adam's rib" strategy

A third way in which single roots can find nearly similar partners to be paired with is to manufacture them out of their own substance. A root (like *chian* "step") may be conjoined with its own cognate accusative construction (like *baoh cumhian* "take a step") to produce a new near-synonym compound *baoh cumhian chian* "step". (Curiously, the order of elements in the resulting compound is fixed: the derived and hence longer form always precedes, systematically violating Behaghel's Gesetz der wachsenden Glieder.)

Baek	*kumneut*	*keut*	"think"
turn	thought	think	
mian cawmneh	*ceh*		"know"
have knowledge	know		
awh sawmnaeuc	*saeuc*		"laugh"
exhaust laughter	laugh		
mian dawmlaj	*tlaj*		"valuable"
have value	valuable		
cia sawmnaen	*saen*		"make an offering"
be offering	make offering		
Praeu tawmbiat	*tbiat*		"embrace, wrap one's arms around"
use embrace	embrace		

Each of these compounding strategies suggests one thing: that form trumps meaning in their construction. The servant word may have a meaning, which is entirely irrelevant (in the cases of "conscription"); there may be a meaningful near-synonym which is then tricked out with a meaningless affix to satisfy the drive for alliteration (in the cases of "prosthesis"), or the original form may be recycled to produce an entirely new partner (the "Adam's rib" examples). All of these strategies suggest the plausibility of the whole cloth strategy, having in common with it the putative drive to create formally symmetrical compounds with little or no semantic motivation.

2.4 Rhyme-swapping

A strong argument against the whole cloth theory may seem to be that there is no productive mechanism for generating the vast bulk of meaningless forms – this in spite of the fact that servant word compounds number in the thousands. Without such a mechanism, it seems that the empty forms must be learned rather than spontaneously generated. But in fact, this brings us to our fourth argument: there is a recognized productive mechanism at hand, one well known to Khmer speakers as the verbal game of making *piak kunloah kat* '(making) inverted words' (There is a similar game in Thai, called *kham phuan* 'flipped words', cf. Iwasaki & Ingkaphirom 2005: 46–7). In English there are a small number of portmanteau words like *smog, motel, brunch* which are generated by the same means: two words in conjunction or apposition swap their rhyme portions: *smoke* + *fog* yields *sm-og (+ *f-oke).* The result of this rhyme swapping is the generation of a meaningful portmanteau (*smog*) plus a nonsensical word (*foke*) which is so completely "discarded" that English speakers have difficulty producing it. Khmer speakers have made an institutionalized game from swaps of this sort. Sometimes, the semantic result of rhyme swapping seems to include a tinge of pejorification: *baaj tmaat* (literally 'rice + vulture') yields *baat tmaaj* (literally yes (respectful, male speaker)' + nonsense form), the meaning of which is something like "Yes, you're saying 'yes' all the time, but your 'yesses' (*Baat*) are bogus (*?tmaaj*)"; *sdaaj aac* (literally 'regret + be able' yields *sdac aaj* (literally 'king' + nonsense form, the meaning of which is "worthless or lousy king".

Sometimes, the rhyme swapping game produces a nonsense word which can be combined with one of the words from which it is formed (almost always, the one with which it alliterates): Chiang (2002: 10) cites the pair *krawmom krawmac* 'young girl', whose first member means 'young girl' and whose second member is a servant word. But *krawmac*, he argues, is the result of the rhyme swapping game applying to the expression

Krawm-om	*l'aaw*	*pd-ac*
'girl	good	exceptional'

which becomes, by a commonly recognized inversion:

krawm-ac	*l'aaw*	*pd-om*
'nonsense word +	good +	nonsense word'

Chiang cites many other cases of this sort, some of them perhaps less compelling than others (to the extent that he makes up some cases of input structures that are not recognized by other speakers). But it seems that the game of making *piak kunloah kat* provides at least one possible mechanism whereby new alliterating servant words can be generated.

2.5 Recursion: Decorative symmetry run wild

Finally, if more evidence is needed that symmetrical conjunction is a hypertrophied indulgence of Khmer, that evidence is provided by the fact that to a mild degree all of these strategies are conjoinable: symmetrical conjunction may be recursive, producing threesomes and quads.

2.5.1 *Synonym + Servant word compound*

laun +	*rauhah*	*rauhuan*	"fast, quick"
quick	fast	{twin form}	
kiak	*keut* +	*ceut*	"close to"
close	{twin form}	close	
c'aet hawl +	*skawp*	*skawl*	"satiated"
full full	full	{twin form}	

2.5.2 *Etymological Doublets + Servant word compound*

panjnjaa paathii+	*praac*	"intelligence"
intelligence {twin form}	intelligence	
(Pali)	(Sanskrit)	

2.5.3 *Synonym + Adam's Rib Compound*

cia	*cumlooh*	*clooh* +	*prawkaek*		"argue"
be	argument	argue	dispute		
bawnjceenj	*paunlww*	*plww* +	*ceunjcaeng*	"illuminate, brighten'"	
emit	light	illuminate	brighten		

The "whole cloth" idea that new forms could originate from original doublets is a well-established finding in other fields like biology (Mayr 2002: 38 et passim). The most common and harmless mutations (whether of genes or of larger structures) are simple replications: A > AA. Through a later series of developments, AA > Aa, and finally Aa > AB. The novel (paralogous) "B" form is finally free to deviate not only in form but in function from the original (orthologous) "A" form of which it was at first a perfect, and then later an imperfect replica. A similar trajectory has been proposed for the development of variations, harmony and counterpoint out of the simultaneous or serial repetition of one original melody in Western music (Lach 1925: 13–4).

3. Non-referential bulking is not pragmatically motivated elsewhere

A number of researchers, from Karlgren 1923 [1962] and Bloomfield (1933: 395–6), through Bolinger (1975: 438) Matisoff (1978: 13, 1982: 74–6, 2001: 295) McGregor (2001: 206) and Heath 1998 have proposed that etymologically unmotivated

phonological bulking or "thickening" may often occur to prevent a word from eroding away completely. For Austroasiatic languages, in fact, Anderson & Zide 2002 have proposed a specific "bimoraic constraint" which is responsible for the addition of various etymologically unmotivated enlargements to "unacceptably short" roots. The possibility that some non-esthetic motivation for bulking may exist in Khmer is quite strong, given the ferocity of phonetic reduction processes in the language (Huffman 1970, passim, Haiman 1998: 612–3, Haiman & Ourn, to appear).

Nevertheless, the native intuition that such bulking characterizes *sawmnuan vauhaa*, "elegant style" should not be disregarded, nor the fact that bulking of the sort that we are looking at is particularly prominent in ritual language of the same sort as the New Year's message. (cf. Malkiel 1973: 354–5 for a rare instance of esthetically motivated phonological bulking in Spanish: *mur ciego* 'bat' literally 'blind mouse' became *murcie-la-go*, the "prime mover" being a "purely esthetic delight in a characteristic syllabic-accentual arrangement").

Nor should we overlook the existence of a variety of related phenomena in languages elsewhere:

3.1 Baby talk/doggerel

Lo llamaba con todas las variaciones de su nombre: Platero! Plateron! Platerillo! Platerete! [She would call [the donkey] by all the wheedlingly affectionate variations of his name: –Platero! Platero, you big goof! Darling little Platero! Cute old Platero! Gooch 1970: 19–20 translating Jimenez 1952: 74–5]

Motherese and/or "doggerel" (pet-language) in a variety of languages is a well known source of etymologically unmotivated reduplication (Paul 1880: 182; Sapir 1921 [1970]: 76; Lach 1925: 17; Kelkar 1967: 48; Schachter et al. 1979: 97). While the conventional thinking is that baby talk registers are exclusively an adult creation (Paul op.cit. 181–2), and that replication is thus possibly motivated by a drive for greater clarity, there is also some evidence that the bisyllabic template for words like *yumyum* and the like originate with infants themselves (Oller 1978, 2000; Stoel-Gammon & Otomo 1986), and that adults are just following their lead.

3.2 Game trash talk

I'm the Dude, or his Dudeness, or El Duderino…

Games are a well-known context for ludic distortions and additions. Consider the card players in Gogol's *Dead Souls*:

> *tšervi!* *tšerv-ototšina*!
> Hearts heart- ???
> *Pik-endras!* *Pi – tšuruštšux !* *pi- tšura ! … .* *Pi- tšuk*!
> Spade -??? Spade – ??? spade – ??? … .. spade – ???
> (Gogol 1842 [1972]: 11)

A standard translation (MacAndrew 1961: 23) gives us only
"Hey, hearts, hearties... Spades, spadies, ..spuds"
but V. Nabokov provides a more nuanced appreciation (Nabokov 1981: 24) :

> *Chervi* means "hearts", but it also sounds very much like "worms", and with the linguistic inclination of Russians to pull a word out to its utmost length for the sake of emotional emphasis, it becomes *chervotochina* which means "worm-eaten core". *Piki* –spades – French *piques* –turn into *pikentia,* that is, assume a jocular dog-Latin ending; or they produce variations as *pikendras* (fale greek ending) or *pichura* (a faint ornithological shade), sometimes magnified as *pichurushchuk* (the bird turning as it were into an antediluvian lizard, thus reversing the order of natural evolution).

Or the trash talking Hungarian "chessplayers" in Karinthy's sketch of that title:

> *[Hol jutnak bele az esz- é be]*
> where they-come into the mind- your- into
> "how do they occur to your mind"
> *a kis esze – mesze-jé- be*
> the small mind ??? your into
> "your little mindy pindy"
> *kis mesze –esze– jé- be*
> small ??? mind your into
> "to your pindy mindy"
> *A kis mesze- esze- vesze- jé –be*
> the small ??? mind ??? your into
> "to your pindy mindy shindy" (Karinthy 1975: 353)

Comparable verbal behavior is noted on the part of Anglo baseball players whose infield chatter has been described by (among others) Dave Barry.

3.3 "Aggressive" reduplication

As noted by long ago Paul (1880: 187) and more recently by Zuraw (2002: 395) there exists a purely phonological drive to make adjacent words and syllables come closer together in sound, the more so if the meanings of these words are imperfectly understood. Thus *smorgasbord* and *hoc est corpus* become *smorgasborg* and *hocus pocus* in English. *Ramadan* and *mocca faux* become *Remmidemmi* and *Muckefuck* in German. Wälchli (2005: 126) suggests that exactly this process is responsible for the phonological convergence of bare binomials.

3.4 Agreement

It is possible that formal grammatical agreement itself may have originated as an outcome of "aggressive reduplication" (cf. Ferguson & Barlow 1988: 17, who do

not hesitate to ascribe its origin to an esthetic drive). Against those who doggedly maintain the functional/referential motivation of grammatical agreement, we should note:

a. the immediate abandonment of agreement in all pidgins.
b. the complete absence of agreement in an impressive number of languages (Paul 1880: 304)
c. the galumphing appearance of agreement from a strictly pragmatic point of view (Dahl 2005: 201) reckons that speakers of Spanish "could save millions of hours of conversation every day, and the average Spanish novel could be twenty to thirty pages shorter if gender markers and agreement were deleted." Compare the well-known scorn of Jespersen who called agreement superfluous, cumbersome, and primitive (Jespersen 1894[1993]: 45, 1924: 207).
d. the completely non-functional appearance of agreement in a variety of languages (Hagège 1993).

While normally agreement may be claimed to serve the ends of tracking reference or marking constituency, this is emphatically not what agreement does in languages like the ones Hagège surveys (ibid. 76–88).

Positive evidence that agreement is an esthetically motivated phenomenon also exists. First is the well-established typological fact that formal agreement is a short-range phenomenon (Comrie 1975; Corbett 1979), hence grossly similar to the "aggressive reduplication" observed by Paul, Wälchli and Zuraw. In the same way that it is only words which are adjacent which are subject to reduplication, it is only words that are adjacent that are subject to formal agreement. As distance between "trigger" and "target" increases, formal agreement is replaced by "semantic" agreement, which is really independently motivated anaphora. A compact example from Bulgarian is the sentence

Vi	*ste*	*bolen*
You (2PL.)	are (2PL.)	sick (3SG.MASC.)

addressed to a single male. The copula agrees in person and number with the polite plural addressee, but the predicate adjective, already too distant for formal agreement to apply, "agrees" only with that addressee's actual gender and number, which are not signaled in the subject pronoun. (The same phenomenon is attested in Romanian, and in the orthography of standard French.)

Second there is the (admittedly weak) diachronic evidence furnished by Slavic languages that agreement was formerly a more recognizable rhyming phenomenon than presently. The short form of the adjective rhymes with the head noun, while the long form, a later development incorporating the definite article, does not (Vaillant 1958: 496).

Third there is the synchronic evidence from languages which are developing agreement systems. Plank 2003 describes his dialect of Bavarian as one which is developing multiple marking of (in)definiteness within a noun phrase, as attested in the following examples (the innovative replica is indicated in capitals):

Was	*ganz*	*WAS*	*Neues*
Something	quite	something	new

'something quite new'

zwei	*ganz*	*ZWEI*	*alte*	*Brezn*
two	very	two	old	pretzels

'two very old pretzels'

EIN	*so*	*ein*	*Depp*
one	such	one	fool

'such a fool'

Whether this becomes established even in Bavarian is unclear, but the reappearance of the same number and gender marking morpheme in more than one place within a NP looks like familiar agreement, using lexical rather than affixal material. As a native speaker, Plank is eminently qualified to address the semantic, pragmatic, or esthetic force of these referentially redundant repetitions. He talks about "spirited emphasis" "a characteristically emotional flavor", and finally, that "it is only natural that peoples rules by passions and given to laissez-faire" should express themselves in this way (375). While not everyone might be satisfied by this description, it is notable that Plank nowhere attempts to motivate the repetition in any referentially functional way.

Matters are even clearer in at least one other language where a novel agreement system may be observed in statu nascendi, Puerto Rican (or possibly, New World) Spanish (Poplack 1980). What she observed in Puerto Rican Spanish, where final [-s] was a sociolinguistic variable, was the variable retention or dropping of this sound in noun phrases like

La-s	*chica-s*	*bonita-s*
The (PL.)	girl (PL.)	pretty (PL.)

A priori: Some version of the principle of least effort might dictate that the final sound is everywhere deleted. Some version of the principle of clarity might dictate that it is never deleted. Some version of competing motivations might predict that it is retained, but only once. Given the presumable validity of both "least effort" and "clarity" we should anticipate either the compromise or (in some version of Optimality theory) the persistent prioritizing of one or the other principle (if unmarkedness trumps faithfulness, then dropping throughout; and vice versa).

What Poplack found instead was that "there was a tendency for concord at the string level" (1980: 377): if the first word dropped the final [-s], so too did the

others; if the first word retained the final [-s], so too did the others. This carefully quantified result is incompatible with any of the a priori predictions above. It is however, fully compatible with a referentially neutral drive for formal symmetry – and, to the extent that functionalism recognizes only the pragmatic motivations of clarity and laziness, it is incompatible with functionalism, as Labov 1994 emphatically makes clear. However, it makes little sense to turn around and attempt to characterize this behavior as "a principle of least effort at the grammatical level" (Labov 1994: 559), since to do so is to vitiate the concept. The PLE promotes dropping phonetic material. If it promotes retaining it, then "it" is not longer the PLE, but something else. We propose that it is the same (ludic or esthetic) drive which is responsible for "aggressive reduplication", and for the decorative symmetry of symmetrical compounds in languages like Khmer.

3.5 Structural priming

The (morpho-) syntactic analog of "aggressive reduplication" is "structural priming": the [semantically and] pragmatically unmotivated tendency to repeat the general syntactic pattern of an utterance" (Bock & Griffin 2000: 177). Indeed both of these modern notions (aggressive reduplication and structural priming) may have been prefigured in traditional discussions of non-iconic analogy via "contamination", the process which makes the words "four" and "five" begin with the same sound, irrespective of their etymological sources (**kwetores* and **penkwe*: e.g., Bloomfield 1933: 409). What is common to all of these devices is the imposition of etymologically and semantically unmotivated similar structure on to different chunks of discourse. As Bock & Griffin (among others) reiterate, "speakers repeat themselves" (op.cit. 177) and their reasons for doing so warrant our respectful inquiry.

4. Conclusion

The preceding remarks are a plea for the reconsideration of what we mean by "style" and by "ritual" in language.

4.1 Style

A tradition exemplified in classical grammars such as Leumann-Hoffmann-Szantyr (1965: 785–90), parodied in Queneau 1947, and one that goes back at least to Aristotle holds that style is the dispensible packaging, while the referential content is the indispensible core of any public behavior – or for that matter, of language. First comes the message/theme, then all the elaborations/variations that constitute

the style. Certainly the bare bones functionalism of von der Gabelentz and his many predecessors and successors recognizes only the two countervailing drives for clarity and the least expenditure of effort in the history of languages. There is nothing we have said to challenge this view of style as a playful unnecessary add-on, one that we find in ritual language more than in ordinary language, in ordinary languages more than in pidgins. Eliminate the flourishes, and you risk sounding "harsh" in SE Asian languages like Khmer. Eliminate grammatical agreement, and you sound ungrammatical, but still comprehensible in pidginized versions of inflecting languages.

Yet if stylistic elaboration is the equivalent of playful "galumphing" (Miller 1973), decorative art in general (Boas 1925), or the outcome of an esthetic drive, then probably "its roots go very deep" (Humphrey 1973 speculates on a cognitive basis for our pleasure in rhymes), and in fact there is no archeological or ethnographic evidence for the historical priority of representation over decoration. It may be that there are other aspects of language structure in which the workings of an esthetic drive are revealed not only in ritual but in everyday language, possibly even in the most stripped down versions of human communication.

4.2 Ritualization

It seems that most discussions of grammaticalization or ritualization in the linguistic literature today are actually discussions of conventionalization in the sense most compactly illustrated by Bellugi and Klima in their well-known 1976 paper on sign language. Iconic charades become the conventional signs of ASL or any other sign language through two parallel processes: on the one hand, standardization and on the other, reduction. The first of these is roughly equivalent to analogy, and the second to grammaticalization plus a great deal of sound change. (Other scholars have suggested that the trajectory from pidgins to creoles is comparable (Givón 1979), or even the trajectory from paralanguage to ordinary language (Fónagy 2000; Bolinger 1975). In an earlier article on ritualization (Haiman 1994), one of the authors of this article did the same. It now seems to us that ritual clearly involves more than mere convention, complex as that is, and that a creative galumphing drive for elaboration needs to be acknowledged, which is separate from both standardization and reduction.

Such a recognition would have implications for attempts to study the origin of language. One reason why historical linguistics is still disjoint from pre-historical linguistics, is that the former seems to provide us with two reliable mechanisms of conventionalization (sound change, analogy), but none whatsoever for the genesis of language: all words are traced back to other words (see Paul 1880 chapter 9 for

a prescient exception). It is as though geological theory recognized only erosion. Armed with a richer understanding of ritualization, we may be closer to proposing a universalist account of language evolution, wherein the processes underway at the present time and accessible to our observation, offer all the data that are needed to account for all phases of the development of language.

References

Anderson, Gregory D.S. & Norman H. Zide. 2002. Issues in proto-Munda and proto-Austroasiatic Nominal derivation: The bimoraic constraint. In *Papers from the Tenth annual meeting of the Southeast Asian linguistics society*, M. Macken (ed.) 55–74. Flagstaff AZ: Arizona State University Press.

Barlow, Michael & Charles Ferguson 1988. Introduction. *Agreement in natural languages*, M. Barlow & C. Ferguson (Eds), 1–22. Stanford CA: CSLI.

Bellugi, Ursula & Edward Klima. 1976. Two faces of the sign. In *Origins and evolution of language and speech,* S. Harnad, H. Steklis, J. Lancaster (Eds), 514–38. New York NY: Academy of Sciences.

Bloomfield, Leonard. 1933[1961]. *Language.* New York NY: Holt.

Boas, Franz. 1925 [1966]. Stylistic aspects of primitive literature. In *Race, language, and culture*, 491–502. Boston MA: Beacon.

Bock, Kathryn & Zenzi, M. Griffin. 2000. The persistence of structural priming. *Journal of Experimental Psychology: General:* 129(2): 177–92.

Bolinger, Dwight. 1975. *Aspects of language*. 2nd Edn. New York NY: Harcourt Brace Jovanovich.

Bright, William. 1990. With one lip, with two lips. *Language* 66(3): 437–52.

Chiang, Caaun. 2002. *Veejaaukaawr Kmaer*. (Khmer grammar.) Phnom Penh: Privately printed.

Comrie, Bernard. 1975. Polite plurals and predicate agreement. *Language* 51: 406–18.

Corbett, Greville. 1979. *Hierarchies, targets, and controllers*. University Park PA: Penn State University Press.

Chuon Naath et al. 1938. *Dictionnaire cambodgien*. Phnom Penh: Institut Bouddhique.

Dahl, Östen. 2005. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Farghal, Mohammed. 1992. Colloquial Jordanian Arabic tautologies. *Journal of Pragmatics* 17: 223–40.

Fónagy, Ivan. 2000. *Languages within language*. Amsterdam: John Benjamins.

Fox, James J. (Ed.) 1988. *To speak in pairs*. Cambridge: CUP.

Geertz, Clifford. 1955. *The religion of Java.* New York NY: Free Press.

Givón, Talmy. 1979. *On understanding grammar*. New York NY: Academic Press.

Gogol, Nicolas. 1842 [1972]. *Mertvyje dushi* (Dead souls). Minsk: Izdatel'stvo Belarus'.

Gooch, Anthony. 1970. *Diminutive, augmentative and pejorative suffixes in Modern Spanish,* 2nd Edn. Oxford: Pergamon Press.

Gorgoniev, Yu. 1966. *Grammatika Khmerskogo jazyka* (Grammar of the Khmer language). Moskva: Izdatel'stvo Nauka.

Gregerson, Kenneth J. 1984. Pharynx symbolism and Rengao phonology. *Lingua* 62: 209–38.

Hagège, Claude. 1993. *The language builder.* Amsterdam: John Benjamins.

Haiman, John. 1994. Ritualization and the development of language. In *Approaches to grammaticalization,* William Pagliuca (Ed.), 3–29. Amsterdam: John Benjamins.

Haiman, John. 1998. Possible origins of infixation in Khmer. *Studies in Language* 22: 597–617.

Haiman, John & Noeurng Ourn. To appear. Creative forces in Khmer. *Papers of the Southeast Asian Linguistics Society* 12.

Heath, J. 1998. Hermit crabs: Formal renewal of morphology by phonologically mediated Affix substitution. *Language* 74(1): 728–59.

Heine, Bernd & Mechtild Reh 1984. *Grammaticalization and reanalysis in African languages*. Hamburg: Helmut Buske.

Hock, Hans-Henrich & Brian Joseph 1996. *Language history, language change, and language Relationship.* Berlin: Mouton-de Gruyter.

Huffman, Franklin. 1970. Modern Spoken Cambodian. New Haven CT: Yale University Press.

Humphrey, N. 1973. The illusion of beauty. *Perception* 2: 429–39.

Iwasaki, Shoichi & Preeya Ingkaphirom. 2005. *A reference grammar of Thai.* Cambridge: CUP.

Jakobson, Roman. 1966 [1971]. Grammatical parallelism and its Russian facet. *Collected Writings of Roman Jakobson.* Vol. 2, 98–135. The Hague: Mouton.

Jespersen, Otto. 1894 [1993] *Progress in language*. Amsterdam: John Benjamins.

Jespersen, Otto. 1924 [1992]. *The philosophy of grammar*. Chicago IL: The University of Chicago Press

Jimenez, Juan Ramon. 1952. *Platero y yo*. 12th Edn. Buenos Aires: Losada.

Karinthy, Frigyes. 1975. A sakkozók (The chess players). *Görbe tükör* (Warped Mirror), 353–5. Budapest: Szépirodalmi Könyvkiadó.

Karlgren, Bernard. 1923 [1962]. *Sound and symbol in Chinese*. London: OUP.

Kelkar, Ashok. 1967. Marathi baby talk. *Word* 20: 40–54.

Khin, Sok. 2002. *La grammaire du Khmer moderne*. Paris: You-Feng.

Labov, William. 1994. *Principles of linguistic change: Internal factors*. Oxford: Blackwell.

Lach, Robert.1925. Das Konstruktionsprinzip der Wiederholung in Musik, Sprache, und Literatur. *Akademie der Wissenschaften in Wien: Philosophisch-historische Klasse, Sitzungsberichte* 201: Band 2: 3–40. Wien: Hölder-Pichler-Tempsky.

Leumann, Manu, Johann B. Hoffmann & Anton Szantyr, A. 1965. *Lateinische Grammatik,* Band II. Munich: Beck.

MacAndrew, Andrew R. 1961. *Dead souls.* (translation of Gogol 1842]. New York NY: Signet.

Malkiel, Yakov. 1973. Genetic analysis of word-formation. In *Current trends in linguistics,* Vol. 3, T. Sebeok (Ed.), 305–64. The Hague: Mouton.

Marchand, Hans. 1960. *Categories and types of English word formation*. Heidelberg: Carl Winter.

Maspero, Georges. 1915. *Grammaire de la language khmère*. Paris: Imprimerie Nationale.

Matisoff, James A. 1978. *Variational semantics in Tibeto-Burman*. Philadelphia PA: Institute for the Study of human institutions.

Matisoff, James A. 1982. *The grammar of Lahu*, 2nd Edn. Berkeley CA: University of California Press.

Matisoff, James A. 2001. Prosodic diffusibility in South-East Asia. In *Areal diffusion and genetic inheritance*, A. Aikhenvald & R.M.W. Dixon (Eds), 291–327. Oxford: OUP.

Mayr, Ernst. 2002. *What evolution is*. New York NY: Basic Books.

McGregor, William. 2001. *A functional grammar of Gooniyandi*. Amsterdam: John Benjamins.
Mikone, Eve. 2001. Ideophones in the Balto-Finnic languages. In *Ideophones,* Erhard F.K. Voeltz & Christa Kilian-Hatz (Eds), 223–33. Amsterdam: John Benjamins.
Miller, Stephen. 1973. Means, ends, and galumphing: Some leitmotifs of play. *American Anthropologist* 75: 87–98.
Moravcsik, Edith. 1978. Reduplicative constructions. In *Universals of human language,* Vol. 3, J. Greenberg (Ed.), 297–334. Stanford CA: Stanford Univeristy Press.
Nabokov, Vladimir. 1981. *Lectures on Russian literature*. F. Bowers (Ed.), New York NY: Harcourt, Brace, Jovanovich.
Nacaskul, Karnchana. 1976. Types of elaboration in some Southeast Asian languages. *Austroasiatic Studies*, Part 2, P. Jenner et al (Eds), 873–90. Honolulu HI: University Press of Hawaii.
Nguyen, Dinh-Hoa. 1965. Parallel constructions in Vietnamese. *Lingua* 15: 125–39.
Oller, D. Kimbrough. 1978. Infant vocalization and the development of speech. *Allied Health and Behavioral Sciences* 1: 523–49.
Oller, D. Kimbrough. 2000. The emergence of the speech capacity. Mahwah NJ: Lawrence Erlbaum Associates.
Ourn, Noeurng & John Haiman. 2000. Symmetrical compounds in Khmer. *Studies in Language* 24: 483–514.
Paul, Herman. 1880 [1995]. *Prinzipien der Sprachgeschichte*. 9e Edn. Tübingen: Niemeyer.
Plank, Frans. 2003. Double articulation. In *Noun phrase structures in the Languages of Europe,* F. Plank (Ed.), 337–95. Berlin: Mouton de Gruyter.
Poplack, Shana. 1980. Deletion and disambiguation in Puerto Rican Spanish. Language 56 (2): 371–85
Queneau, Robert. 1947. *Exercices de style*. Paris: Gallimard.
Ratliff, Martha. 1992. *Meaningful tone: A study of tonal morphology in compounds, form classes, and expressive phrases in White Hmong*. Northern Illinois University Center for Southeast Asian Studies, Report No. 27.
Rischel, Jørgen. 1995. *Minor Mlabri*. Copenhagen: Museum Tusculaneum.
Roffe, G. Edward. 1975. Rhyme, reduplication, etc. in Lao. *Studies in Tai linguistics in honor of William Gedney,* 285–317. Central Institute of English Language.
Sapir, Edward. 1921 [1970]. *Language.* London: Harvest.
Schachter, Frances Fuch et al. 1979. *Everyday mother talk to toddlers*. New York NY: Academic Press.
Smyth, David. 2002. *Thai: An essential grammar*. London: Routledge.
Stanford, James. 2007. Sui adjective reduplication as poetic Morpho-phonology. *Journal of East Asian Linguistics* 16(2): 87–111.
Stanford, James. To appear. Lexicon and description of Sui adjective intensifiers. *Linguistic Discovery* 4(2)
Stoel-Gammon, Carol & Kiyoshi Otomo. 1986. Babbling development of hearing-impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders* 51: 33–41.
Svantesson, Jan-Olof. 1983. *Kammu phonology and morphology*. Malmo: CWK Gleerup.
Ts'ao Ts'ui-yün. 1961. A preliminary study of descriptive words in the Miao language of Eastern Kweischow. In *Miao and Yao linguistic studies: Selcetd articles in Chinese, translated by Chang Yü-hung and Chu Kwo-ray* [Linguistic Series V, Data paper no. 88, Southeast Asia Program Cornell University], H. Purnell, Jr. (Ed.), 187–210. Itaca NY: Southeast Asia Program, Cornell University.

Vaillant, André. 1958. *Grammaire comparée des langues slaves*. Lyon: IAC.
Vongvipanond, Peansiri. 1992. Lexicological significance of semantic doublets in Thai. In *Papers on Tai languages, linguistics, and literatures in honor of William J. Gedney on his 77th birthday* [University of Michigan Monograph on Southeast Asia, Occastional Paper 16], C. Compton & J. Hartmann (Eds).
Wälchli, Bernhard. 2005. *Co-compounds and natural coordination*. Oxford: OUP.
Weidert, Alfons. 1973. *I Tkong Amwi: Deskriptive Analyse eines Wardialekts des Khasi.* Wiesbaden: Otto Harrassowitz.
Wierzbicka, Anna. 1987. Boys will be boys. *Language* 63(1): 95–114.
Zuraw, Kie. 2002. Aggressive reduplication. *Phonology* 19: 395–439.

Time management formulaic expressions in English and Thai

Shoichi Iwasaki
University of California, Los Angeles

Abstract

Time management formulaic expressions are linguistics resources that provide the temporal edge for the speaker who needs to speak and think at the same time. Prefabricated complement taking constructions ('it seems to be') and the partially prefabricated pseudo-cleft ('what they did was') are such expressions. Compared to more established time buying expressions ("let me just say") and discourse markers ("y'know'), time management formulaic expressions are still in the process of coming to be associated with the function of buying time. Through repeated use, these expressions may acquire a stronger tie with the specific function of buying time. In addition to the time management expressions in English, this paper will also discuss the case of the topic construction in Thai.

0. Introduction

The literature of formulaic expressions often refers to their linguistic and sociolinguistic advantages (Fillmore 1979; Coulmas 1981; Pawley & Syder 1983b, Pawley 1985;

Aijmer 1996; Matisoff 1979; Wray & Perkins 2000 among others). Linguistically, they help create native-like texture; "Nativelike fluency… is probably unattainable until the language learner has memorized a repertoire of formulae" (Pawley 1985: 90). Sociolinguistically, they help create contextual appropriateness; "[Formulaic expressions offer] the comfortable feeling that one is saying exactly the right thing under a particular set of circumstances" (Matisoff 1979: xxiv). Often noted as well is their cognitive advantage; "[Formulae] can be drawn from the memory without much effort, and at the same time, they give us time for conversational planning (Coulmas 1981: 9–10)," "[they free] speakers from concentration on the mechanics of speech production – the tasks of finding and articulating words for their thoughts and ensuring that utterances are grammatical and idiomatic" (Pawley 1985: 92, see also Ochs 1979: 73).

This paper is concerned with the cognitive-temporal advantage that formulaic expressions bring about for the speaker who must think and speak simultaneously. Though present in all formulaic expressions, there are special formulae for which the primary task is to make time for the speaker to organize discourse and to prepare the listener to receive upcoming information. Wray & Perkins (2000) use the term 'turn holders' (e.g., "And another thing; let me just say") and 'fillers' (e.g., "If the truth be told; if you want my opinion; if you like") to refer to these types of formulaic expressions.[1] I will use the term "time management formula" (TMF) to explore a similar but different type of expression. Unlike Wray & Perkins (2000)'s 'turn holders' and 'fillers,' the TMFs that I examine are not (yet) specialized formulaic expressions, but they are showing signs of forging a link with such a function.

More specifically, in this paper I will focus on the TMFs that take the form of complement-taking expressions and the pseudo-cleft in English and the topic construction in Thai. The TMFs that take the form of complement-taking expressions are in many respect prefabs (e.g., "it seems to be"), and many appear with the first person singular subject and a complement-taking predicate in the present tense ("I think," "I remember"). These are likely to be "stored and retrieved whole from memory at the time of use" (Wray & Perkins 2000: 1). The pseudo-cleft is a partially prefabricated expression with a restricted range of verbs and the copula form ("what they did was"); in most cases, it takes the form of *what (NP)* {*do/happen/say*} {*is/was*} (Hopper 2004: 4). The Thai topic construction is the least prefabricated; it is a grammatical construction containing an open slot followed

1. Turn holders and fillers are two types of "time-buyers" identified by Wray & Perkins (2000) – Note that their use of the term 'filler' is different from the use adopted in this paper, as will become evident shortly. Other time-buyers they identify are 'discourse shape markers' ("There are three points I want to make") and 'repetition of preceding talk' ("(A: What's the capital of Peru?") – B: What's the capital of Peru? (Lima, isn't it?)).

by the so-called topic marker /nîa/. However, I consider this to be a formulaic expression which can satisfy a set of criteria for TMFs to be presented later.

To analyze these formulaic expressions, I propose to adopt a dynamic-bidirectional model for the form-and-function mapping in language based on the general tenet of the emergent model of language (Du Bois 1985; Hopper 1987, 1988, 1998). In this model, both processes of a particular linguistic structure (e.g., "I think") coming to take on a particular function (e.g., gaining the temporal edge) and that of a function (e.g., a need to gain the temporal edge) coming to find a particular structures (e.g., "I think") to be realized are considered two integrated aspects that bring about linguistic changes. Through repeated occurrence of these processes, a progressively strengthened tie between form and function will emerge. Methodologically, research of the dynamic-bidirectional model starts by noticing speakers' needs in communication and linguistic resources they employ to accommodate them in actual context, then seeks a possible reason why such resources are suitable to address the communicative demand.

The traditional functionalist framework works well for a grammatical structure that is transparently related to a certain function. For example, the left-dislocation in English is a particular structure that performs the function of showing contrastive topic. The need to call attention to the contrastive topic motivates a marked word order with that topic articulated with a prominent accent at the beginning of a sentence. However, dynamic-bidirectional model is more suitable to analyze a form-function relationship that is currently emerging, and has not been clearly brought to the analysts' attention. The dynamic-bidirectional model can help us identify a range of structures that are employed for performing the task of buying time and, at the same time, help us understand why such structures are chosen to do the task.

Section 1 introduces the concept of 'idea/image transfer' during speech production, a cognitive process crucial for understanding TMFs. In Section 2, after I compare TMFs with fillers, hedges and discourse markers and provide a working definition of TMFs, I examine complement-taking expressions and the pseudo-cleft in English as examples of TMF. In Section 3, I turn to Thai and examine a particular formula involving the so-called topic marking particle, /nîa/. In Section 4, I suggest that different grammars develop for spoken and written languages (the multiple-grammar model), and TMFs are a quintessential grammatical form cultivated for spoken language.

1. 'Idea/image transfer' and 'time management formulaic expression (TMF)'

During the process of linguistic communication, abstract ideas and pre-linguistic mental images emerge and dissipate constantly within speakers' consciousness

(Chafe 1979). Language provides a means to transform such pre-linguistic entities of various types and sizes into a string of linearly ordered linguistic units. In this study this process is referred to as 'idea/image transfer.'

The pre-linguistic idea/image is a particular form of consciousness, which is created by memory and perception (Yamadori 2002: 14).[2] An idea/image may be as concrete as a scene with a clear agent, patient and other types of participants (e.g., a witness of a traffic accident), or as abstract as an idea with various degrees of complexity (e.g., the effect of minus ions in the future). Ideas/images are arranged and combined to further create various mental activities; awareness of objects, actions, emotions, and the formation of opinions, attitudes, desires, and decisions (cf. Chafe 1994: 31).

Speaking is, in part, the process of conveying these mental activities into linguistic forms on-line. It is the process of "selecting the gist or gists of thousands or millions of not necessarily conscious ideas to be transformed into a particular linguistic expressions of those ideas" (Schank 1990: 26). In other words, the "speaker … is engaged in a real-time process of focusing on a sequence of ideas and converting these ideas, one after another, into language" (Chafe 1979: 166).

A smooth idea/image transfer is one important concern for speakers in communication. Upon reflection on our own communication processes, we realize that idea/image transfer is a constant, on-going process throughout a communicative event. (This is particularly evident when we try to express difficult ideas; when we have something to say, but can't express it in words.) We can also be aware that the transfer process can and often does happen while speaking. Speakers can manage idea/image transfer successfully with relative ease for certain types of ideas/images, but not for others.[3] In this latter situation, rather than simply resorting to silence, the speaker may exhibit such linguistic behaviors as "backing and filling, trying out phrases and discarding them often before they have been completed, revising and expanding phrases already uttered, adding phrases as afterthoughts to sentences already completed, and so on. (Chafe 1979: 167)."

2. What is perceived is called 'perceptual mental image' and what is stored is 'stored mental image' in neuro-psychological literature. When a perceptual image can be matched with a stored mental image, "understanding" will take place. (Yamadori 2002: 27–32).

3. The degree of difficulty of idea/image transfer is to some extent related to the availability of lexical and phrasal resources. For example, to express a complex psychological experience such as 'anger' there are such ready-made phrases (i.e., formulaic expressions) as 'he made me mad,' 'his words really hurt me.' 'he is so rude.' etc.

During such a phase, the speaker also produces audible noises to buy time and to signal that they are working on an idea/image transfer.[4] These noises range from simple pre-linguistic vocalizations ('uhm'), hedging expressions ('like') to discourse markers ('I mean,' 'you know' etc.). 'Time management formulaic expressions' (TMFs) are similar to these devices, but while these devices are a dedicated means for buying time, TMFs are a more indirect, but resourceful, device which not only gives speakers a temporal edge, but also provides the addressee some clues regarding the information that is being prepared for presentation.

2. Time managing formulae in English

In this section, I will examine how complement-taking expressions work as a time-management formula in English. Before examining TMEs, however, we need first to observe how speakers produce fluent speech, and second, how they employ fillers, hedges and discourse markers to buy time for the organization of thought while speaking.

2.1 Fillers, hedges and discourse markers in English

The idea/image transfer is hidden underneath speech, not directly observable in most cases when the process goes without a hitch. The next excerpt shows such a successful case. Here, the speaker, Patrick, a male undergraduate student, is describing his experience of the major earthquake that hit Los Angeles in 1994.[5] He successfully transferred the awareness of his action (lines 1, 3), emotion (line 2), and perceptions (lines 5 through 9) into words. The only signs of a possible transfer problem are seen in the discourse marker 'well' (line 2), the one-second pause (line 4), and the hedging expression 'like' (line 6). (The truncation in line 9 is a slip of the tongue, and is a more micro level disfluency rather than the macro level disfluency we are concerned with.)

4. One can argue that disfluency results from a speaker's concern about how to present information in a socially appropriate manner. This is true, but I maintain that such effects from communication are a type of cognitive burden, which results in disfluency.

5. The English data, unless otherwise noted, come from my Northridge Earthquake Conversation Database. The English data consist of 4 dyadic conversations between UCLA students. Two students previously unacquainted with each other were invited at the time to discuss their experiences of the Northridge earthquake in 1994. The data was collected by myself and transcribed by Kathy Howard approximately one year after the earthquake. Names used in the transcripts are all pseudonyms.

(1) EQ#

1.	when I came out of the building,	[ACTION]
2.	well first of all I was in shock.	[EMOTION]
3.	but when I came out of the building,	[ACTION]
4.	… (1.0)	
5.	I could smell gas,	[PERCEPTION]
6.	and I could see fires like,	[PERCEPTION]
7.	down a block or two?	[PERCEPTION]
8.	they had a lo: t of buildings condemned.	[PERCEPTION]
9.	on third- on thirteenth.	[PERCEPTION]

The process of idea/image transfer surfaces for inspection when a speaker encounters a problem with this process. In such a situation, he/she often uses fillers and hedging expressions to cope with the problem. Consider the following excerpt of a conversation between two students, Jake (a male undergraduate student) and Rosa (a female graduate student). Jake is trying to transfer his opinion into words by first stating that he liked the Bay Area (line 1) because "it's really cool" (line 3). He tried to elaborate on the reason, but he stumbled because a potentially contradictory idea arose in his consciousness, i.e., Oakland (part of the Bay Area) is actually not cool because it is dangerous. He verbalized this problematic idea in lines 11–12.

(2) EEQ #2

1.	J:	I really like the Bay-	[OPINION]
2.		*uh like* the Bay Area.	
3.		I think it's really cool.	[OPINION]
4.		[*uhm*],	
5.	R:	[So do] I.	[OPINION]
6.	J:	*I mean like,*	
7.		*I don't know*,	
8.		*I like*,	
9.		*I mean it's kinda' like*,	
10.		(.5)	
11.		Oakland's sorta' dangerous,	[OPINION]
12.		but.	
13.	R:	Depending on where y-	[OPINION]
14.		you know,	
15.		where you go.	

Jake's verbal behavior, especially the lines between 6 and 9 and the half-second pause in line 10, are a reflection of his struggle to construct and transform pre-linguistic ideas (OPINION, his attitude toward a city) into language forms on-line. Discourse markers ("I mean," "I don't know,") and a hedging expression

("it's kinda like") buy time for the speaker, though they did not help the speaker in this case.

The next is an excerpt from a different conversation from the same data set between two male undergraduate students. This time, discourse markers and hedging expressions ("I mean," "you know," and "like") assisted the speaker to a successful transfer of his mental image (ATTITUDE toward an earthquake).

(3) EEQ4

1.	just, (.2) just from what I've experienced.	[EXPERIENCE]
2.	*mean* I've experienced *like*,	[EXPERIENCE]
3.	*you know like*. (.) five earthquakes.	[EXPERIENCE]
4.	*like* I remember,	
5.	*and you know*,	
6.	*I mean* I always get a kick out of 'em.	[ATTITUDE]
7.	*you know?*	
8.	*I mean* I always think they're kind of fun?	[ATTITUDE]

As seen here, the speakers rely on fillers, hedges and discourse markers when they need time for the idea/image transfer. However, they may also employ more subtle resources of time-management formulae, as seen in the remainder of the paper.

2.2 A working definition of time-management formulaic expressions

In this section, I first examine a couple of putative time-management formulaic expressions, and then, through a comparison of these with fillers, hedges and discourse markers, I provide a working definition for them in order to further identify similar formulae. We start with the excerpt reported in Pawley & Syder (1983a: 201), paying particular attention to lines 1 and 5. (A dash indicates the location of a pause. Transcription conventions have been slightly simplified.)

(4) Pawley & Syder (1983: 201)

1.	→	*and it seems to be* –
2.		if a word is fairly – – high on the frequency list –
3.		I haven't made any count -
4.		but – just – – impressionistically – – um
5.	→	*um – the chances are* –
6.		that you get a – compound –
7.		or – another – – phonologically deviant – – form –
8.		with ah which is already in other words
9.		which is fairly frequent – ly the same – phonological shape –

Regarding this segment Pawley and Syder make the following comment; "[the speaker] has not planned the actual word content of his discourse very far ahead. The pattern of his dysfluencies (frequent pauses and reformulations – SI) indicate (sic.)

that he does this planning a few words at a time" (202).[6] However, this speaker's strategy is not to rely on hedges and discourse markers alone, but also on such expressions as "it seems to be –" (line 1) and "the chances are -" (line 5).

Notice in particular that "it seems to be –" (line 1) is a complement-taking expression but no complement clause appears after it. Rather than constructing a complex sentence with this clause, the speaker produced an unrelated conditional clause in line 2. Here, the speaker seems to have executed a complement-taking expression "prematurely," before he organized his thoughts into a well-formed complement clause. Rather than dismissing this use of complement-taking expressions as ungrammatical, however, I propose that it is becoming a formulaic expression of time management (TMF) as part of the grammar of spoken language. To elaborate on this point: the speaker can throw in a prefabricated expression such as "it seems to be" to create a framework for upcoming discourse, thereby gaining time to formulate the exact wording to be used next. In other words, from the perspective of the dynamic-bidirectional model of form-function mapping, the speaker needing to buy time finds complement taking structures useful. As a consequence, these structures are becoming more strongly associated with the function of buying time for discourse organization.

Though the other complement taking expression "the chances are –" (line 5) is followed by a *that*-complement, its meaning is very weak. In fact, this expression can be completely omitted and a consequence clause can simply be supplied for the conditional clause from line 2 ("(if a word is fairly high on the frequency list), you often get a compound"). This shows that the speaker may be trying to halt the speed of sentence production for the rather complex ideas expressed on lines 7–9.

To summarize the discussion so far: I first introduced the process of the transfer of abstract ideas and pre-linguistic mental images into linguistic forms. Though this process cannot be directly observed, the traces of such effort can be seen clearly in the use of fillers, hedges, and discourse markers when speakers have a problem with the process. I proposed that complement-taking expressions work as time management formulae (TMF) as well, but unlike fillers and the like, they also stamp the speaker's stance for the ensuing discourse. We will return to this last point in the next section.

It is useful at this point to clarify differences between time-management formulae, on the one hand, and fillers, hedges, and discourse markers, on the

6. Indeed Pawley & Syder (1983a: 200) contrast this transcript with another of a similar length containing many instances of "y'know" and fillers such as "um" and "ah"; more precisely in this other transcript they found 7 instances of "y'know" and 8 instances of fillers within a 17 line transcript.

other. Based on the brief discussion to follow, I will give a working definition of time-management formula. Three properties, two structural and one functional, will be discussed. First, fillers, hedges, and discourse markers take more stereotypic forms, and thus are easily identifiable as such. Time management formulae, on the other hand, take a more inconspicuous form, like complement-taking expressions, and can blend in the discourse as normal looking expressions, and in that sense they are not completely specialized. Second, the time management formulae are more constrained in terms of discourse positions. This contrasts with discourse markers, which may appear at boundaries of various units in discourse; "tone groups, sentences, actions, verses and so on" (Schiffrin 1987: 36). Fillers and hedges appear even more liberally. The time management formulae, on the other hand, appear at the discourse periphery, usually at the beginning, of a discourse segment.

Third, meanings of discourse markers such as "I mean/Y'know" (Schiffrin 1987) have become very abstract, and their ubiquitous appearance in informal conversation is an attestation of their semantically less substantial nature. In contrast, the semantic content of time-management formulae is strong, making it possible to indicate a particular type of framework (e.g., stance marking) that is opened for the subsequent discourse.

I summarize these points as parts of a working definition of time-management formulaic expressions. With this definition, other similar formulae can also be identified.

A working definition: Time management formulaic (TMF) expressions

Formal properties

1. TMFs take more elaborate forms than fillers, hedges and discourse markers
2. TMFs appear at the beginning of a discourse segment

Functional property

3. TMFs, because of (1) above, provide a specific framework for upcoming information

2.3 Complement-taking predicates and TMF

In Section 2.2, we observed that complement-taking expressions such as "it seems to be" are used for TMF, but this process can be observed on a much more general level. Complement taking expressions consisting of a subject and a "Complement-Taking Predicate (CTP)" (Noonan 1985) turn out to be particularly useful constructions for the purpose of buying time while creating a framework

for further development of discourse. First, English places the "main" clause (e.g., CTP clause) before a "subordinate" clause, and further it allows the "main" clause to appear more or less independently from the "subordinate" clause. According to Thompson (2002), many CTP clauses in conversation are executed separately from the complement clause; CTP clauses can appear after a complement clause (e.g., "because she uh= has had enough **I guess**." (143)), or without any associated clauses (e.g., ".. this is = , … pepsin, I think, .. **I'm not sure**," (144)). If the speaker produces such fragmentary CTP clauses first, it will give him/her time before uttering the main thrust of the idea that he/she wishes to express. In other words, the speaker gains time for discourse organization by using a CTP. The more CTPs are used in this way, the stronger the link between the CTP and the time buying function becomes.

Second, the CTP can announce the framework before making a core statement in the "subordinate" clause. The most frequently occurring CTPs in Thompson (2002)'s database are "think/thought, know/knew, guess, remember."[7] Thompson notes that these "most frequent, and therefore the great majority, of CTPs in the data… (provide) epistemic, evidential, or evaluative *framing* for the utterances they go with." (141, emphasis added). The findings in Diessel and Thomasselo (2001)'s study of the acquisition of complement clauses by young children are consistent with this view in that most CTPs[8] in child language data "function as epistemic markers, attention getters, or markers of the illocutionary force" (132).

Relative independence of CTPs and the framing function provide an impetus for CTP to create the temporal management function, which is to announce the speaker's stance for a discourse segment that will follow, while giving the speaker time to organize wording of that discourse segment.

In what follows, I analyze in detail the expression "I remember" (one of the most frequent complement taking verbs found in Thompson's study above[9]) as an examples of TMFs. In the next excerpt, Pete is trying to explain why he is not afraid of earthquakes. In line 4, he utters "I remember."

7. These tend to appear as formulae with the first person pronoun and without a complementizer. In addition to these predicates, 'see/saw' appear quite often, but behave differently from these four in that the subject of 'see/saw' is more varied and tends to appear with a complementizer.

8. They use the term CTV (complement taking verbs) in their study.

9. It is interesting to note that "remember" is used differently in Diessel and Tomasselo's child language acquisition data. In the child language data, most uses are related to the expression "Do you remember?" to indicate that "the associated preposition conveys information that is familiar to the interlocutors due to shared experience" (121). In the adult data we are examining, "remember" is often used in the prefabricated formula "I remember," and refers to the speaker's own memory.

(5) EEQ#2

	1	F:	So you're not afraid of earthquakes at all?
	2	P:	no,
	3		no.
→	4		*I remember,*
	5		I mean,
	6		when they're,
	7		I remember walking to school,
	8		when I was a kid,
	9		(.7) in the-,
	10		what.
	11		(1.5)
	12		I don't remember what,
	13		what earthquake that was.
	14		But it was a good size one,
	15		it was like a six,
	16		or seven.
	17		=I lived in Glendale at the time,
	18		(.4) uhm.
	19		(.4) But I remember watchin' the sidewalk.
	20		shimmy.
	21		and [shake.]
	22	F:	[was that] about seven years ago?

After Pete answered Fred's question ("So you're not afraid of earthquakes at all?") with "no, no", he started to assemble information from a memory concerning his experience with another earthquake during his childhood. He prefaced his past experience story with the complement-taking expression, "I remember." This expression qualifies as a time managing formula because it appears at the beginning of a discourse segment; it takes a more elaborate form than fillers and the like, and because of this, it gives a particular framework for the upcoming discourse (i.e., "I am going to tell you my past experience story"). That is, the expression 'I remember' indicates that the process of idea/image transfer has begun in a certain direction, while giving the speaker time to organize mental images for linguistic presentation. Notice even after the speaker gave himself time with "I remember," he added the discourse marker "I mean" to gain more time. And he continued to struggle to construct his memory in language during almost his entire turn.

From the perspective of the dynamic-bidirectional model, time management formulae are not completely fixed with the function of buying time. It is being negotiated in the current state of the English language in general as well as in the specific excerpts we are examining in this paper. While the speaker is aware that he/she can buy time with more fixed formulae such as "If you want my opinion;

There are three points I want to make" (Wray & Perkins 2000: 16), he/she may not be completely aware of the effect that 'I remember' and other CTPs mentioned earlier can have. In many cases, 'I remember' is not deliberately chosen to obtain time for discourse organization, but simply ends up giving the speaker such a temporal advantage. The point to underscore here is that the forms that TMFs take are not intrinsically designed for the specific purpose of buying time, but they are in the right shape to carry out this function. In other words, the need to buy time finds an opportunistic site in the complement taking expressions.

In the next excerpt, Pete again uses the same formula ("I remember") when he assembles information from memory. A one second pause follows, but Fred did not barge in because "I remember" makes Pete's intention to continue speaking clear, thereby holding the floor for Pete. The self-directed question, "how old was I," is also a time managing formula of a different kind, which specifically reveals a problem in the cognitive process.

(6) EEQ#2

	P:	well that happens every time.
		every time,
		there's some,
		some prediction.
		(.4)[like],
	F:	[yeah].
→	P:	I remember,
		(1.0)
	P:	hhhhh well how old was I.
		I guess I was like in high school,
	F:	mm hmm,
	P:	and- and,
		it came around,
		there-
		there was some Nostradamus [prediction],

In his study of "remember," Tao (2001) found 399 tokens of "remember" in the 1.5 million word corpus. Interestingly, nearly half of the "remember" tokens appear with the first person subject, and close to 40% of all the tokens appear without any materials following. In other words, the form "I remember" as shown in the two excerpts above is a typical form in which "remember" appears in spoken discourse.[10]

10. The functions that Tao found for this formulaic expression are epistemic (certainty and uncertainty) and meta-linguistic (e.g., "attention getting, "tying" and "soliciting addressee responses"). Tao's statistics also show about one third of first person subjects with 'remember' are in the negative, in the form of 'I don't remember.'

2.4 Pseudo-cleft and TMF

The pseudo-cleft construction also works similarly to the complement-taking expression (cf. Pawley & Syder 1983b, Hopper 2000, 2004). Hopper claims that a pseudo-cleft is a prefabricated expression (2004: 4) based on his finding that 90% of all the pseudo-clefts in his spoken discourse corpus are constructed with an extremely small number of verb types, in the form of *what (NP) {do/happen/say} {is/was}*. He further found that what follows the *wh-* clause is usually not a single NP as intuition based analysis of the pseudo-cleft often assumes (e.g., "*what you need is a new suit*"), but a stretch of discourse that extends to more than a single clause. The next exerpt clearly exhibits these points (2000: 5). The brackets are added to indicate the clefted element with a *wh-* word.

(7) Hopper (2000: 5)
and then [what they've done is] the cellar it's got a cellar and you go down the steps to the cellar but there's like a proper two proper rooms so on your left you've got a sort of cellar with a quarry tile

In this stretch of discourse, we find no single grammatical constituent that semantically corresponds to *what they've done*. That is, although *what they've done* anticipates according to a prescriptive grammar 'a set of actions of the builders/decorators,' what actually follows is 'a description of the interior' (Hopper 2000: 10). In other words, the speaker has begun the idea/image transfer with this pseudo-cleft, but has not designed the entire sentence.

Some pseudo-clefts such as *what I {suppose/suggest} is* or *what happens is* are semantically empty (Hopper 2000: 13), suggesting that their function is getting close to what discourse markers perform, i.e., assisting the on-line production of spoken discourse. Based on these observations, Hopper lists the following as pseudo-clefts' functions:

a. To alert the listener that an upcoming utterance is noteworthy
b. To make an attitudinal comment on an upcoming utterance
c. To state a general theme for the upcoming utterance
d. To buy time while alternative wordings are considered
e. To hold the floor pending the upcoming utterance

Further examination of pseudo-clefts has lead Hopper (2004) to suggest that they may be classified with other expressions (time-management formulae in our terminology) such as *the point is/the idea is/the thing is*. These constructions including the pseudo-cleft are exactly like other complement-taking expressions we have examined earlier (*I remember, it seems to be, the chances are*) in that they all involve the initial announcement component followed by the main thrust of important information.

2.5 An interim summary

We have uncovered various means that are available for English speakers when they need extra time to prepare a transfer of abstract ideas and pre-linguistic images into language forms. We have examined complement-taking expressions and pseudo-clefts. They are useful because they provide the speaker with the needed temporal edge and announce the framework for the upcoming discourse without revealing that he is using the time for discourse organization.

The process of idea/image transfer is expected to be a cognitive universal relevant in communication contexts, and speakers in any linguistic community use some linguistic means to address this problem. However, it is possible that different languages may use different linguistic resources in order to achieve the same goal. In the next section, we see a drastically different linguistic resource that is recruited in Thai conversation for the same purpose in spoken language communication.

3. Time managing formulae in Thai

Like English speakers, Thai speakers use various linguistic means to manage problems that may arise when transferring abstract ideas and pre-linguistic images into words. And similar to English speakers, they take advantage of particular grammatical structures to cope with this cognitive-linguistic demand. In this section, we will first examine fillers, hedges and discourse markers in Thai, and then examine in detail one particular time management formula, /[open slot] + nîa/, a formulaic construction composed of a slot for a lexical, phrasal and clausal element followed by /nîa/. (Iwasaki 2008).

3.1 Fillers, hedges and discourse markers

The following is a list of tokens used frequently by Thai speakers to signal problems of idea/image transfer. Some have grammatical uses or clear lexical meanings, but as fillers, hedges, and discourse markers, they provide time for on-line processing like their English counterparts. Long vowels in the words shown below are often shortened in informal speech.

/ʔəə, mm/etc.	fillers (pre-linguistic vocalization)
/khɯɯ/	a hedging expression, used also as a copula
/bèɛp/	a hedging expression (lit. 'type' 'kind'), similar to English 'like'
/kɔ̂ɔ/	a hedging expression/discourse marker, also used as a linking particle (LP) with the meanings of 'also, because, so'; also used as part of conjunctives like /lɛ́ɛw-kɔ̂ɔ/ 'and then'
/wâa/expressions	discourse markers, e.g., /lîak-wâa, mǎythŭŋ-wâa, khɯɯ-bèp-wâa/ etc. similar to 'in other words, that is… , what I mean is..'; /wâa/ is also used as a verb of 'saying' and a complementizer; used as part of conjunctives like /prɔ́-wâa/ 'because'
/kɔ̂ɔ-bèɛp, kɔ̂ɔ-bèɛp-wâa/	multiword hedging expressions composed of the elements above

In the next excerpt,[11] Tip reported the verbal exchange that she had with her friend on the phone. Both Tip and her friend were Los Angeles residents at the time of the earthquake, but the friend was out of town on the day when the earthquake hit. Her friend had called Tip to find out the extent of the earthquake damage. Tip produced discourse rather fluently.[12]

(8) TEQ#3[13]

511 Tip: kháw mây dây (.) kháw mây dây yùu thîi-nîi ŋay
3 NEG get 3 NEG get stay here PP
'She wasn't home, you see.'

kháw kɔ̂ɔ thoo maa thăam wâa
3 so phone come ask COMP
'So she called me.'

pen-ŋay-mâŋ ʔalay
how. is.it what
'(She asked) how it was.'

11. The Thai data come from my Northridge Earthquake Conversation Database (collected by myself and transcribed by Amy Meepoe Baron) and my other collections. See footnote 5 for a description of the Northridge Earthquake Conversation Database.

12. In the second line, Tip used /kɔ̂ɔ/, but this use is more grammatical ('so') than hesitational.

13. Abbreviations used for Thai transcriptions: 1 (first person), 2 (second person), 3 (third person), ASP (aspect), CAUS (causative), CLS (classifier), CM (challengeable information marker), COMP (complementizer), COP (copula), DM (discourse marker), HDG (hedge), HES (hesitation), LINK (linker), LP (linking particle), NEG (negative), NIMP (negative imperative), PP (pragmatic particle), REC (reciprocal), QP (question particle), SLP (speech level particle), YS (young sibling), .hhh (audible in-breath).

bɔ̀k mây-pen-lay
tell OK
'(I told her) everything is OK.'

kháw bɔ̀k ca kàp maa
3 tell CM return come
'She said she'd come home.'

bɔ̀k yàa kàp
tell NIMP return
'(I told her) not to come back.'

yàa phûŋ kàp ma lɔ̌ɔy
NIMP soon return come PP
'Don't come back yet!'

hây (.) yùu toŋ-nán lè dii lɛ́ɛw
CAUS stay there PP good ASP
'It's better to stay there.'

thɔ̌ɔ chôok dii
you luck good
'You are lucky'

thîi thɔ̌ɔ mây yùu bâan thɔ̌ɔ
COMP you NEG stay house you
'that you didn't stay at your house.'

Speaker Ae in the next excerpt, on the other hand, used many tokens of hesitation markers; /kɔ̂-bèp, bèp-wâa/.

(9) TEQ#4

708 Ae: → kɔ̂-bèp khít yùu nay cay (.)
HDG think stay in heart
'Like (I was) thinking in my heart'

→ ʔé nî man: bèp-wâa imɛjineshân lŭ khwaamciŋ (.)
what this it HDG COMP imagination or reality
'if this is like my imagination or reality'

lɛ́w man kɔ̂ sân pay tâŋ:: ʔathít kwàa
and it LP shake go all week more
'and it kept shaking more than a week'

→ lɛ́w kɔ̂ tham-hây law bép-wâa
and LP CAUS we HDG
'and it made us like'

→ tháŋ ʔathít kwàa nî bèp-wâa lúsŭk (.)
all week more tghis HDG feel
'(It lasted) the whole week, I felt.'

tch! aníi sîi talɔ̀t welaa
(click) this PP always time
'"Oh, this is it!", (I said to myself) all the time'

phɔ́-wâa man sàn lɛ́-kɔ̂
because it shake and
'because it shook and'

→ hhh .hhh lɛ́w-kɔ̂ [bèp-wâa]
and HDG
'and'

3.2 Time-management formula: The /__ nîa/construction

In addition to fillers, hedges and discourse markers, Thai speakers use time management formulae like English speakers, though the forms they use often are radically different. In this section, we are concerned with the formula, /__ nîa/, a contraction of the proximal demonstrative /níi/and the pragmatic particle /na/. While the expression /nîa/is often described as a topic marking particle (Ekniyom 1982; Diller & Juntanamalaga 1989; Iwasaki & Ingkaphirom 2005),[14] speakers take advantage of a function associated with topic presentation and use the expression to buy time for idea/image transfer. This function is important in spoken discourse, but not in written text. This point will be elaborated on later in the discussion section.

According to the working definition given in Section 2.2, /__ nîa/qualifies as a time management formulaic expression. It will become clear how this formulaic expression provides the speaker time to develop discourse on-line. Compared to the fixed form of fillers, hedging expressions and discourse markers, the /__ nîa/ formula takes a more elaborate form in that it has an open slot before /nîa/in which a variety of linguistic expressions (a word, phrase, and clause) can appear. Compared to the English TMFs considered earlier, this Thai formula is much more flexible with almost complete freedom for the open slot.

14. Besides /nîa/, other demonstratives are also used as a 'topic marker'. /nán/and /níi/(both with the high tone) are a proximal and distal demonstrative, respectively, used in formal style. /nîi/(with the falling tone) is a proximal demonstrative, and like /nîa/is used in informal style. /nîa/is the only marker that is created by fusing a demonstrative and a pragmatic particle.

The /__nîa/expression in (10) below works as a time managing formula. It provides time for the speaker to formulate thoughts on-line while producing information incrementally. It also serves to indicate that the process of conveying a coherent idea/image has just begun.

(10) Ads

1 A: tua mitsubichi nîa na há
CLS Mitsubishi NIA PP SLP
'Mitsubishi, you see'

2 sămràp tua phŏm lέɛw nîa
as.for CLS 1 ASP NIA
'for me, you see'

3 yîihɔ̂ ɯ̀ɯn lέɛw nîa
brand other ASP NIA
'(if compared with) other brands, you see'

4 B: khà
'yes'

5 A: mây dây nɯ̌a kwàa yìihɔ̂ɔ ɯ̀ɯn
NEG ASP superior than brand other
'(it) is not really superior to others.'

The idea that the speaker intends to convey is 'In my opinion Mitsubishi is not superior to any other brands.' In line 1, the speaker starts with the formula, 'Mitsubishi /nîa/' followed by /na/(pragmatic particle) and /há/(speech level particle). This use of/nîa/represents a typical topic marking function (see more discussion later), but as can be seen in the transcript, the topic noun, 'Mitsubishi,' is not directly followed by a comment in the next line. Instead, lines 2 and 3 add further topic-like phrases ('for me' and 'other brands').

Although /__ nîa/is a useful time management formula for a speaker, it does not always guarantee successful delivery of utterances. Thus a /nîa/-marked noun phrase prematurely executed may not be successfully closed with a comment. The speaker in (11) produces two /nîa/marked NPs in succession. First she produces 'the ability of the (VISA) card /nîa/,' and then 'the people who can use it /nîa/'. The second noun phrase is followed by the comment provided in line 305 ('are those in the high class of the society'), but the first one is abandoned. This shows that the speaker had not formulated the entire utterance when he produced the /nîa/phrase in line 303. Notice that although the speaker failed to complete the utterance at first, he was successful in announcing the general direction of talk by giving a temporary framework, and in keeping his floor while working on image transfer.

(11)

303 lέɛw-kɔ̂ khwaam-săamâat khɔ̌ɔŋ bàt nîa
and ability of card NIA
'and the ability of the (VISA) card, you see'

304 khon radàp thîi cà cháy dây nîa
people rank COMP CM use can NIA
'the people who can use it, you see'

305 kɔ̂-khɯɯ khon thîîi radàp sŭuŋ khûŋ pay
are people COMP rank high go-up go
'are those in the high class of the society.'

The two examples examined above show how /__ nîa/provides time for the speaker to formulate complex information through incremental presentation.

The next excerpt demonstrates that a speaker sometime uses /__ nîa/for the sake of the addressee. In (12), the interviewer asks one of the interviewees a rather complex question: 'What message do you think the commercial you like is trying to communicate to us?' Since the interviewer is likely to have prepared a question, producing a complex sentence is possible. In contrast, processing such information may go beyond the cognitive capacity of the addressee. In fact, the addressee did not understand the question at first and requested that it be repeated. Then, in the utterance quoted below, the interviewer used /nîa/three times to break up this complex information: 'the commercial you like,' 'Oreo (commercial),' and 'after watching the commercial.'

(12) Ads 365

Nisa: thăam wâa ¦ ə̀ə nɔ́ɔŋ khít wâa ¦ khôotsànaa thîi nɔ́ɔŋ chɔ̂ɔp nîa ¦
ask COMP ¦ HES YS.2 think COMP ¦ commercial COMP YS.2 like NIA ¦
'I asked … uhm.. What do you think the commercial that you like, you know,'

ooliioo nîa ¦
Oreo NIA
'the Oreo commercial, you know,'

duu lέɛw nîa ¦
look ASP NIA
'after watching it, you know,'

kháw tôŋkaan cà bɔ̀ɔk alay kàp law
3 want CM tell what with 1
'What do (you think) they want to tell us?'

In this section, I demonstrated that a speaker uses the formulaic expression /__ nîa/to gain time and present part of the complete information when he has not

formulated, or cannot formulate, a sentence in advance. A speaker also provides manageable chunks of information with this formulaic expression when dealing with a complex utterance, so that the listener can process smaller pieces of information, bit by bit.[15]

3.3 "Challengeable" information and /nîa/

The /__ nîa/formula which we investigated above performs the same function of buying time as English complement-taking expressions and the pseudo-cleft. However, this is made possible through a different route than the English counterparts. We will examine different functions of /nîa/in different contexts, and suggest that the concept of "challengeability" (Givón 1982b) is crucial to understanding the function of time management associated with this formulaic expression.

It has already been noted that /nîa/is often described as a topic marker. The next excerpt shows this function of /__ nîa/clearly. This excerpt is from a transcript of a parent-teacher conference at a college in Bangkok. In line 67, a female teacher (FT) marks /khanɛɛn/'grade' with /nîa/. This leftmost NP ('grade') represents a given/identifiable concept, which was first mentioned in line 65, and is followed by a comment, 'reaching 1.5 GPA.' All these are necessary properties for a topic NP.

(13) Parent-Teacher

65 FT: ʔà lɛ́w kɔ̂ khanɛɛn kɔ̂ mây dii ná khá
HES LINK LP grade LP NEG good PP SLP
'And her grades are not good, you know.'

66 P: lɔ̌ kháp
'Is that right?'

67 FT: khanɛɛn nîa ¦ thɯ̌ŋ nɯŋ cùt hâa lɯ́- plàaw ¦ kɔ̂ mây sâap
grade PP ¦ reach one point five QP ¦ LP NEG know
'About her GPA, did it reach one point five or not ? I don't know. (I am not sure if her GPA last year was at least one point five.)

15. The strategy performed by /__ nîa/to create a discourse is not unique to Thai. Tao (1996) shows that Mandarin speakers produce a series of NP intonation units which describe different aspects of a concept before producing a comment. English speakers use the so-called 'try markers' (e.g., "Remember Tom?") before forming a complete sentence (Sacks & Schegloff 1979; Keenan & Schieffelin 1976). What distinguishes the Thai case from these examples is that Thai has developed the specialized expression /__ nîa/to perform this task more automatically. In this respect, the /__ nîa/may be closer to the case reported by Chafe (1976) on Caddo (Oklahoma); speakers of this language use an quotative particle (e.g., it-is-said) after a 'premature' topic.

While the above description of 'grade' as a topic seems reasonable, one question arises. When a referent is given information in Thai discourse, it normally does not appear overtly (zero anaphora). Therefore, we must ask why some given information is restated. Here, the notion of 'challengeable' information as proposed by Givón (1982a, b) is helpful. That is, although the referent, 'the grade' has been mentioned in the discourse, its saliency in the mind of the addressee at the time of utterance 67 in (13) can be questioned (i.e., challengeable). It is in such a case that the speaker restates given information in order to prepare the addressee for an upcoming comment. In other words, /nîa/is more than a simple topic marker.

The concept of challengeability is a probabilistic scalar notion that replaces the logician's binary notions of presupposition and assertion. It is the speaker who determines the degree of challengeability in a given communicative context for any information. Non-challengeable information is shared information, but this information can be shared in various ways.

> "The speaker assumes that a proposition **p** is *familiar* to the hearer, likely to be *believed* by the hearer, *accessible* to the hearer, within the reach of the hearer etc. *on whatever grounds*." (1982b: 100 – italics in the original)

Any information outside this range is challengeable to various degrees. While non-challengeable information can be expressed without any evidential justification, challengeable information requires support with evidential marking. I suggest here that though /nîa/may not be an evidential morpheme per se, it is a marker that acknowledges that information is potentially challengeable.

Marking challengeable information is extended to conditional clauses and questions. The sentence in (14) below consists of the conditional clause 'if you drive a Mitsubishi' and the subsequent clause 'you will become like those in the commercial – I mean – you will be happy.' Notice that the conditional clause is marked with /thâa/'if' at the beginning and /nîa/at the end.

(14) Ads

378 thâa khàp mitsubishi khɔ̌ŋ kháw lɛ́ɛw nîa
if drive Mitsubishi of they ASP NIA
'If you drive a Mitsubishi,'

379 kɔ̂ɔ ca pen mɯ̌an nay khôosanaa khɯɯ mii khwam-sùk
LP CM COP same in commercial LINK have happiness
'you will be like those in the commercial – I mean – you will be happy.'

The connection between topics and conditionals has been noted by Haiman (1978). He argues that a conditional clause is a request for the addressee to "accept for a time a proposition *p* which provisionally becomes the framework of reference for the discourse – in particular, for the consequent proposition *q*." (580).

This statement acknowledges that a conditional clause is challengeable.[16] So like the overt restated topic, the conditional clause presents challengeable information and /nîa/marks it.

/nîa/also appears in questions. Questions are challengeable in the sense that the speaker and listener do not share the information. In the next excerpt, Daw is directly quoting what her doctor said after she told him that she took certain medicine. The doctor quoted here disapproved of this prescription, and said 'Which hospital did you go to?'

(15) 311 Daw: khun pay hăa rooŋbaan năy nîa
you go seek hospital which NIA
'Which hospital did you go ?'

We found many questions with /nîa/in a conversation where two students met for the first time. Some examples are shown in (16) below.

(16) EQ#2

197 A: meecâə ʔalay há nîa 'What is your major?'
major what SLP NIA

866 A: mi phî-nɔ́ɔŋ kì khon há nîa 'How many siblings do you have?'
have siblings how.may people SLP NIA

897 A: ʔayú thâwlày lɛ́w há nîa; 'How old are you?'
age how.much ASP SLP NIA

943 A: dĭaw tôŋ mii lian máy há nîa 'Do you a class now?'
soon must have study QP SLP NIA

3.4 An interim summary

Although /nîa/is often called a topic marker, this is only one of several related functions. We discussed earlier that the core function of /nîa/is to mark challengeable information. For the speaker who needs time to organize thoughts into linguistic expression, /nîa/is a convenient resource for gaining time while laying out challengeable information. From the perspective of the dynamic-bidirectional model for form-and-function mapping in language adopted in this paper, we suggest that the connection between the expression and the function of time management will become stronger as the speaker uses it for this function more frequently.

16. Givon (1982b:101-102) demonstrates that some conditionals are challengeable while other are not.

4. Conclusion

In this paper we examined time-management formulaic expressions as realized in complement-taking expressions and pseudo-cleft in English and the /__ nîa/expression in Thai. They are not prefabricated formulaic expressions because they allow freedom within the constructions. However, speakers find these constructions useful for performing the specific function of buying time. As the expressions are used more regularly with this function, they become more formulaic with the connection between the form and the function strengthened (i.e., the dynamic-bidirectional model). At present the expressions examined here are not completely specialized for the function of buying time, but is emerging. English discourse markers such as "I mean" and "y'know" are good examples of expressions that have established a firmer relation between the form and the specific function of buying time. While it is not clear at this point whether other complement taking expressions (e.g., 'I remember') have become specialized like these, it is clear that many complement taking expressions perform multiple functions (Boye & Harder 2007), and this is a common state of affairs for any grammatical constructions that are emerging and exhibit the pattern of "layering" (Hopper & Traugott 1993: 124–6).

Formulaic expressions are useful for the speakers in conversation and other unplanned spoken discourse when they are faced with the challenge of simultaneously processing thoughts and delivering it in language forms. Since normal writing is free of the kind of time pressure that spoken language is subject to, it is natural that these formulae are not employed in written language. In fact, though completely natural in spoken language, complement-taking expressions without a complement as seen in some English examples would be judged ungrammatical and the /__ nîa/would be judged stylistically inappropriate in written text.

This fact, if viewed from the written language's perspective, shows that spoken language is filled with ungrammatical and irregular constructions. However, if viewed from the perspective of spoken language, it is the written language that is deviant. Neither view is correct (Pawley & Syder 1983b).

I suggested elsewhere (Iwasaki, in preparation) that this type of situation is an attestation of language users' possession of (at least) two distinct grammars (the multiple-grammar model). In this model, partially overlapping but different grammars are developed by responding to the specific needs of particular environments of language use – spoken and written modes in this current situation. This ability, similar to the ability of multilingual speakers, is shared by language users in linguistic communities with a literate tradition and allows them to access the most appropriate grammar depending on the context and to activate necessary lexical and grammatical resources.

Due to the limited space I have here, I can provide only one example (pseudo-clefts) to demonstrate the multiple-grammar model. According to Hopper, pseudo-clefts used in conversation are strongly prefabricated constructions with a very limited type of verbs appearing in them; three verbs 'do/happen/say' accounting for almost 90% of all cases (Hopper 2004). Such an almost prefabricated construction is easy to use to gain the temporal edge needed for conversation.

In contrast, written pseudo-clefts use diverse varieties of verbs. This is likely to be related to the fact that written pseudo-clefts do not need to perform the function of buying time and holding the floor, but rather are used to organize complex discourse by making additional and/or contrastive statements to what has come before. Below is a typical example from a newspaper editorial.

(17) (Honolulu Advertiser January 10, 2007)

> "The latter obligation will require many billions of American dollars in physical reconstruction and humanitarian relief of a society shattered by our reckless diversion from the real war on terrorism.
>
> What that obligation does not require, however, is the continued and now-augmented sacrifice of American lives. Such will be the most deplorable price tag of Bush's apparent intention to feed into the Iraq calamity a "surge" of more U.S. troops that will only offer more targets in the sectarian crossfire."

We assume that this editorial writer would use 'pseudo-clefts' differently in his editorials and in his informal speech. Otherwise, his informal conversation would sound too bookish, and his editorial would deteriorate into non-professional, incomprehensible texts.

In this paper, we examined time-management formulae in English and Thai using the dynamic-bidirectional model for the form- and- funtion mapping. I hope to have demonstrated that these formulae are motivated by the special need in spontaneous speech. Our understanding of spoken language grammar is far behind that of written language grammar. We will need to examine other formulaic expressions in spontaneous speech in order to fill this gap.

References

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

Boye, Kasper & Peter Harder. 2007. Complement-taking predicates: Usage and linguistics structure. *Studies in Language* 31(3): 569–606.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and Topic*, C. Li (Ed.), 25–56. New York NY: Academic Press.

Chafe, Wallace. 1979. The flow of thought and the flow of language. In *Syntax and semantics: Discourse and syntax,* T. Givón (Ed.), 159–181. New York NY: Academic Press.

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

Coulmas, Florian. 1981. *Conversational routine: Explorations in standardized communication situations and prepatterned speech*. The Hague: Mouton.

Diessel, Holger & Michael Tomasello. 2001. The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics*. 12–2, 97–141.

Diller, Anthony & Preecha Juntanamalaga 1989. Deictic derivation in Thai. In *Prosodic analysis and Asian linguistics: To honour R.K. Sprigg* [Pacific Linguistics Series C, no. 104], D. Bradley, E.J.A. Henderson & M. Mazaudon (Eds), 169–196. Canberra: The Australian National University.

Du Bois, John. 1985. Competing motivations. In *Iconicity in syntax*. John Haiman (Ed.), 343–365. Amsterdam: John Benjamins.

Ekniyomn, Peansiri 1982. A study of information structuring in Thai sentences. Ph.D. dissertation,The University of Hawaii.

Fillmore, Charles. 1979. On fluency. In *Individual differences in language ability and language behavior,* C.J. Fillmore, D. Kempler & W.S.-Y. Wang (Eds), New York NY: Academic Press.

Givón, Talmy. 1982a. Evidentialilty and epistemic space. *Studies in Language* 6(1): 23–49.

Givón, Talmy. 1982b. Logic vs. pragmatics, with human languages as the referee: Toward an empirically viable epistemology. *Journal of Pragmatics* 6: 81–133.

Haiman, John. 1978. Conditionals are topics. *Language* 543: 564–589.

Hopper, Paul. 1987. Emergent grammar. In *BLS 13,* J. Aske, N. Beery, L. Michaelis & H. Filip (Eds), 139–155.

Hopper, Paul. 1988. Emergent grammar and the a priori grammar. In *Linguistics in context: Connecting observation and understanding,* D. Tannen (Ed.), 117–134. Norwood NJ: Ablex.

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

Hopper, Paul. 2000. Grammatical constructions and their discourse origins: Prototype or family resemblance? In *Applied cognitive linguistics: Theory, acquisition and language pedagogy*, M. Pütz & S. Niemeier (Eds), Berlin: Mouton de Gruyter.

Hopper, Paul. 2004. The openness of grammatical constructions. *CLS* 40: 239–56.

Hopper, Paul & Elizabeth Traugott. 1993. *Grammaticalization.* Cambridge: CUP.

Iwasaki, Shoichi & Preeya Ingkaphirom. 2005. *A reference grammar of Thai*. Cambridge : CUP.

Iwasaki, Shoichi. 2008. What is/nîa/doing/nîa/? : Grammaticalization of topic in Thai. *SEALS XIV. Papers of the 14th annual meeting of the Southeast Asian Linguistics Society 2004*, W. Khanittanan & P. Sidwell, 181–192.Canberra: Pacific Linguistics.

Iwasaki, Shoichi. In preparation. A multiple-grammar model for speakers' linguistic knowledge. Ms.

Keenan, Elinor Ochs & Bambi B. Schieffelin. 1976. Topic as a discourse notion: A study of topic in the conversations of children and adults. In *Subject and Topic*, C. Li. (ed.) New York NY: Academic Press.

Matisoff, James. 1979. *Blessings, curses, hopes, and fears : Psycho-ostensive expressions in Yiddish.* Philadelphia PA : Institute for the Study of Human Issues.

Noonan, Michael. 1985. Complementation. In *Language typology and syntactic description,* Vol. II, T. Schopen (Ed.), 42–139. Cambridge: CUP.

Ochs, E. 1979. Planned and unplanned discourse. In *Discourse and syntax. Syntax and semantics,* Vol 12. T. Givón (Ed.), 51–79. New York NY: Academic Press.
Pawley, Andrew. 1985. On speech formulas and linguistic competence. *Language Modernas* (Chile) 12: 84–104.
Pawley, Andrew & Frances Hodgetts Syder. 1983a. Two puzzles for linguistic theory: Native-like selsection and nativelike fluency. In *Language and Communication*, J.C. Richards & R.W. Schmidt (Eds), 191–227. Longman: Longman.
Pawley, Andrew & Frances Hodgetts Syder. 1983b. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7(5): 551–579.
Sacks, Harvey & Emanuel A. Schegloff. 1979. Two preferences in the organization of reference to persons in conversation and their interaction. In *Everyday language: Studies in ethnomethodology*, G. Psathas (Ed.), 15–21. New York NY: Irvington.
Schank, Roger C. 1990. *Tell me a story: A new look at real and artificial memory*. New York NY: Charles Scribner's Sons.
Schiffrin, Deborah. 1987. *Discourse markers.* Cambridge: CUP.
Tao, Hongyin 1996. *Units in Mandarin conversation: Prosody, discourse, and grammar*. Amsterdam: John Benjamins.
Tao, Hongyin. 2001. Discovering the usual with corpora: The case of *remember*. In *Corpus linguistics in North America: Selections from the 1999 symposium,* R. Simpson & J. Swales (Eds), 116–144. Ann Arbor MI: The University of Michigan.
Thompson, Sandra A. 2002. 'Object complements' and conversation towards a realistic account. *Studies in Language* 26(1): 125–164.
Wray, Alison & Michael Perkins. 2000. The functions of formulaic language: An integrated model. *Language and Communication* 20: 1–28.
Yamadorai, Atsushi. 2002. '*Wakaru' to wa dou iu koto ka – ninshiki no noo-kakgaku* ('What is 'understanding'? – the brain science of cognition.) Tokyo: Chikuma Shoboo.

Routinized uses of the first person expression *for me* in conversational discourse

Joanne Scheibman
Old Dominion University

Abstract

This paper reports on the distributional and functional properties of conversational tokens of the English prepositional phrase *for me*. In these data, *for me* occurs within larger

constructions, and the majority of these uses have pragmatic functions. In its most frequent context following copular predicates, *for me* marks an evaluating speaker (*It's like really hard* ***for me****)*. The expression also appears as a component of a formulaic construction used to make polite requests (*Could you recall those* ***for me*** *please)*. This examination of *for me* illustrates ways in which formal properties of routinized expressions are sensitive to their recurring contexts of use, and also provides a view of how constructions with first person elements fulfill conventionalized interactive functions in discourse.

1. Introduction to the study

Many scholars have noted that formulaic expressions are tied to cultural concepts and recurring social practices (Coulmas 1979, 1981; Pawley & Syder 1983; Aijmer 1996; Overstreet & Yule 2002; Wray 2002). A group of expressions that has been found to be particularly open to fulfilling conventionalized interactive functions are constructions containing first person singular markers (Thompson & Mulac 1991; Dahl 2000; Scheibman 2000; Tao 2001; Scheibman 2002; Kärkkäinen 2003; Ono & Thompson 2003). The relative high frequency of first person elements in discourse in many languages, coupled with their indexical and reflexive character, gives constructions containing these forms expressive and interactional force in discourse contexts.

Characteristic of first person constructions in discourse (and of conversational routines more generally; Coulmas 1981 & Aijmer 1996) is that they often convey meanings beyond those that are strictly semantic or grammatical. For example, in English conversations the expression *I think* does not typically function as a main clause that designates a speaker's cognitive activity; instead the entire expression functions as an epistemic or evidential marker (Thompson and Mulac 1991; Kärkkäinen 2003). Similarly, studies have suggested that in conversations *I don't know* more often functions pragmatically – e.g., to mediate disagreement in conversations – than it does semantically – to convey a speaker's lack of knowledge of a given proposition (Tsui 1991; Scheibman 2000).

In the case of the prepositional phrase *for me*, the expression conveys both grammatical and pragmatic meanings, depending on the construction in which it occurs. *For me* is a grammatical adjunct when it marks a first person beneficiary, e.g., *so he does that* ***for me*** *and that helps me a lot*. More frequently in the conversational data used for this study, however, the expression appears in constructions that have pragmatic functions. For example, the most ritualized use of *for me* is as a formulaic component in polite directives and requests, e.g., *could you recall those* ***for me***. The majority of utterances with *for me* in the corpus, however, are copular constructions in which *me* marks an evaluating speaker, e.g., *it's like really hard* ***for me***, and these expressions are used metapragmatically.

This study, then, examines the formal and functional patterns characteristic of these three contexts in which conversational usages of *for me* occur: (1) as a grammatical adjunct with lexical verbs marking first person benefactive arguments, (2) as a component of a formulaic construction used to form directives and indirect requests, and (3) as an evaluative adjunct appearing with copular predicates. In each of these linguistic contexts, *for me* occurs as part of larger constructions with distinct discourse functions.

2. Routinized expressions ~ routine functions

In his influential paper, *Emergent grammar*, Paul Hopper (1987: 145) proposes that in discourse "grammar is what results when formulas are re-arranged, or dismantled and re-assembled, in different ways". In this view, formal structures are sensitive to, and occur in support of, social and interactional routines. Linguists working within recent usage-based frameworks share the perspective that conventionalization and representation of linguistic structure develop from frequent experiences of use and have shown that communicative and cognitive factors shape linguistic patterning (Hopper 1987; Barlow & Kemmer 2000; Bybee & Hopper 2001; Bybee 2006; Bybee & Eddington 2006). The present study is usage-based in the sense that identification of formal and pragmatic properties of *for me* in conversational contexts is intended both to describe its distribution, as well as to exemplify how routinized expressions are formally and functionally related to their contexts of use.

There is a growing collection of research, notably usage-based studies carried out by Sandra Thompson and her colleagues, which have demonstrated that there are robust links between constructions and their interactional functions in conversational contexts (e.g., Thompson 2002; Thompson & Couper-Kuhlen 2005; Fox & Thompson 2007). Thompson & Couper-Kuhlen (2005: 482) suggest that studies that take as basic the idea that grammar and interaction are mutually contingent promote "the recognition that routinized patterns that we call grammar exist because speakers need routinized ways to implement *actions* [emphasis in original]".

This idea that routinized expressions are tied to interactive and cultural activities has also been prominent in the literature on formulaic language. Pawley & Syder (1983: 209) suggest that a lexicalized element is "a conventional label for a conventional concept, a culturally standardized designation (term) for a socially recognized conceptual category". Fillmore (1977: 25) proposes that formulaic

expressions are indexical in the sense that "their appearance is predictive of a number of details of the situations of their performance". That is, interpretation of indexical elements requires understanding of specific information related to the expression's conventional contexts of use. For example, Fillmore (1977: 24) points out that in using and understanding the expression *next of kin*, speakers know that the person whose relatives are being referred to has to be recently dead, information that is learned from the contexts in which *next of kin* is used.

Coulmas (1979, 1981) discusses *routine formulae* and *conversational routines*, which are fixed expressions tightly linked to particular social situations that carry social meaning (e.g., greetings). Aijmer (1996: 200–201) examines routines such as thanking, apologies, requests and offers, in addition to discourse markers. In a study of clause-final *and everything* and the constructions in which it occurs (e.g., X *and everything, but* Y), Overstreet & Yule (2002: 792) propose that these conventionalized expressions fulfill metapragmatic functions in discourse; namely they are used by participants to clarify contexts in which expectations are intersubjectively acknowledged, which "allows speakers/writers to then offer a justification for thinking contrary to those expectations". This brief review illustrates various ways in which routinized expressions are tied to social contexts.

3. *For me* as a first person singular expression

3.1 Studies of first person elements and prepatterned expression

Relevant to *for me's* conventionalized uses in discourse is the presence of a first person pronoun. As a group, first person singular expressions have been shown to participate in distinct distributional and constructional patterns in interactive discourse (e.g., Dahl 2000; Scheibman 2002; Aaron & Cacoullos 2005). Because constructions containing first person elements make explicit reference to the speaker, they commonly perform indexical functions (e.g., interactional, metalinguistic). Managing relational activities such as negotiating stance, expressing politeness, or performing speech acts are prototypical activities in conversations, and it is not unusual for these functions to lend themselves to prepatterned expression. (Coulmas 1981; Aijmer 1996). Indeed, some first person expressions used in interactive and procedural tasks are so frequent that they have achieved the status of discourse markers (e.g., *I mean, I think,* and *I don't know* in English).

In Schiffrin's (1987) analysis of uses of *I mean* in conversations, she suggests that the combination of *mean* with the first person singular pronoun gives the entire expression metalinguistic functions. Because *I mean* marks speakers' orientations in conversations, "it always has interactional relevance" (Schiffrin 1987: 305). Accompanying *I mean*'s metalinguistic functions is high frequency in discourse. In a study of distributions of subjects, tense, and verb types in English conversations I found that the expression *I mean* accounts for *mean* being the fifth most frequent lexical verb in a database of over 2,100 utterances (Scheibman 2002).

Thompson & Mulac (1991) demonstrate that first person singular subjects with the predicates *think* and *guess* have grammaticized as epistemic parentheticals, and Tao (2001) and Kärkkäinen (2003) also find frequent epistemic uses of the expressions *I don't remember* and *I think*, respectively. In a study of the negative auxiliary *don't*, I found that the auxiliary consistently appears in a reduced form when it is part of the expression *I don't know*, and in the majority of these pragmatic uses the entire construction functions as a politeness marker (Scheibman 2000; also Bybee & Scheibman 1999). In an analysis of Swedish and English conversations, Dahl (2000) reports that utterances with first person singular subjects are highly frequent in conversation, and they show different distributions with respect to lexical verb type than do those with third person subjects. And, finally, in a study of first person pronouns in Japanese conversations, Ono & Thompson (2003) show that first person singular elements more frequently perform nonreferential tasks (e.g., emotive and frame-setting functions) than referential ones. These studies illustrate the conventional social and pragmatic functions of first person expressions in discourse. The dual referential role of *I* or *me* – as both propositional and speech act participant – is characteristic of first person expressions and has consequences for interaction.

3.2. Distribution of *for me* compared to FOR + NON-FIRST PERSON SINGULAR PRONOUNS

Relative to the first person constructions with *I* discussed in section 3.1 (*I think*, *I mean*, and *I don't know*), *for me* is not a high frequency expression. However, when *for me* is compared to formally similar items – that is, combinations of *for* with non-first person singular personal pronouns (*for you, for her/him, for them,* and *for us*) – its distribution is unique. Table 1 displays distributions of 171 prepositional phrases composed of *for* plus human, nongeneric personal pronouns collected from the *Santa Barbara Corpus of Spoken American English*. In this data set, *for me* accounts for 36 percent of the tokens (62/171), which is the largest group.

Table 1. Tokens of *for* +PERSONAL PRONOUN (nongeneric uses) (n=171)

	USES	% TOTAL	LEXICAL VERBS	COPULAR VERBS
for me	62	36%	40%	40%
for you	39	23%	71%	16%
for her/him	37	22%	82%	9%
for them	22	13%	41%	18%
for us	11	6%	50%	0
TOTAL	171	100%		

The sum of percentages in LEXICAL VERB and COPULAR VERB columns do not add up to 100%. Not displayed in these two columns are percentages of utterances without verbs, e.g., only 80% of *for me* tokens occur with codable predicates.

Given in the two right-most columns in Table 1 are percentages of tokens of prepositional phrases appearing with lexical and copular verbs, illustrated with *for me* in (1) and (2), respectively.

1. PREDICATE FINAL *FOR ME* WITH LEXICAL VERB
   they do all that *for me*, (*SBC* 53)

2. PREDICATE FINAL *FOR ME* WITH COPULAR VERB
   you know, it's like really hard *for me*, (*SBC* 04)

When *for me* occurs with a lexical verb, it is more likely that the prepositional phrase will function to mark the pronoun as a benefactive argument, as in (1). However, with copular predicates, the *for*-phrase is typically a component of an evaluative utterance as in (2). This distinction between predicate-final *for me* occurring with lexical versus copular predicates provides an analytically useful device for classifying functions of these expressions using formal criteria and is used to organize analyses in sections 6, 7, and 8.

Note, too, in Table 1 that 40 percent of occurrences of *for me* occur with copular predicates, which is more than double the percentage of *for*-phrases with the next most frequent categories, *for them* (18 percent) and *for you* (16 percent). This more frequent occurrence of *for me* with copular verbs suggests that functions of *for me* are distinct from the other prepositional phrases, specifically that *me* more commonly occurs with *for* in these evaluative constructions than do other personal pronouns.

4. Data and methods

4.1 Data sources

The 395 tokens of utterances with *for me* used for this study were taken from four published corpora of American English spoken discourse, listed in (3). Conversational

excerpts appearing as numbered examples in this paper are tagged with the corpus abbreviation and a number locating them in a conversation (e.g., *SBC* 07) or in the database constructed for this study (e.g., *SWB* 183).

(3)
- *Santa Barbara Corpus of Spoken American English* (face-to-face conversations)(*SBC*), 62 tokens
- *Michigan Corpus of Spoken Academic English* (interactional encounters in academic contexts, e.g., office hours, lab and discussion sessions) (*MICASE*), 101 tokens
- *Switchboard Corpus* (telephone conversations) (*SWB*), 226 tokens[1]
- *AMEX corpus of human-human air travel planning dialogs* (telephone conversations between travel agents and commercial customers) (*AMEX*), 6 tokens

While all four corpora include spoken interactive discourse, there are differences. The majority of the conversations in the *Santa Barbara Corpus* are face-to-face conversations among friends and family members, whereas the *MICASE* data are interactions related to academic activities. The *Switchboard* corpus is unique in that conversations are conducted on predetermined topics between strangers on the telephone. There is evidence that genre differences affect the distribution of *for me* in some cases, e.g., there are proportionally more fronted tokens of *for me* in the *Switchboard* data than in the *Santa Barbara* and *MICASE* corpora (see (4) and (5c) for examples of these constructions). These differences, however, are not central to the analyses in this paper and will not be pursued here.

The corpora also vary in their transcription systems. When possible, examples in this paper are presented using a simplified version of the Du Bois et al. (1993) transcription system, which is the system used to transcribe the *Santa Barbara Corpus of Spoken American English*.[2]

1. Because of the large number of *for me* expressions found in the Switchboard Corpus, only half of the conversations in this corpus were used for the study.

2. Du Bois et al. (1993) transcription symbols used in this paper:

Transitional continuity (final)	.
Transitional continuity (continuing)	,
Speech overlap	[ ]
Pause	..
Truncated	-

4.2 Classification and coding of data

In preparation for analysis, all utterances containing *for me* were entered into a spreadsheet application and coded for a variety of formal, semantic, and pragmatic features. Also recorded for each utterance was information about the speech event (e.g., TYPE OF ACTIVITY) and the speaker (e.g., SEX). Because conversational utterances do not always fall neatly into traditional grammatical classes (Hopper 1997; Scheibman 2002), classification of discourse data is a dynamic process in the sense that coding categories can be added, modified, or even abandoned during different stages of analysis. The most important analytical categories for this study were: MAIN VERB (e.g., *do, be, buy*), PREDICATE ADJECTIVE (e.g., *hard*, *good*), GRAMMATICAL CLASSIFICATION (e.g., adjunct, complement of adjective, independent of clause), FUNCTION (e.g., to mark beneficiary or recipient, focus/contrast), GRAMMATICAL SUBJECT, and EVALUATIVE FEATURES. Using these categories, it was possible to examine distributions and to identify larger constructions in which *for me* occurs.

5. Classifying *for me*

5.1 How "grammatical" is *for me*?

In usage, *for me* is not an expression that is easy to classify. It is distinct from other multiword sequences, or prefabs, (e.g., *in the end, the thing is*) because it is composed of two grammatical elements. Similar to other prepositional phrases in English, the syntactic roles and functions of *for me* vary in discourse.

As discussed in section 3.2, *for me* appears less frequently as an argument of the verb than it does as an evaluative adjunct. Table 2 shows a comparison of the expressions *for me* and *to me* by predicate type. In these samples, *to me* frequently occurs with lexical verbs (87 percent of the tokens) where it marks arguments in the clause, in particular with verbs of communication (e.g., *talk, explain*), sensory verbs (e.g., *seem, look, sound*), and verbs of transfer and motion (e.g., *give, send*). On the other hand, in only 13 percent of the utterances does *to me* occur as an evaluative adjunct with copular predicates (e.g., *it's interesting to me; that's important to me*).

Table 2. Distribution of *for me* and *to me* in conversational data based on predicate types (two different samples)

	LEXICAL VERBS		COPULAR VERBS		TOTAL	
for me	70	32%	147	68%	217	100%
to me	86	87%	13	13%	99	100%

In contrast with *to me's* distribution, in this data set over two thirds of *for me* tokens occur with copular verbs (68 percent), and only 32 percent with lexical verbs. *To me*, then, more commonly marks recipient and experiencer arguments of the verb than *for me* marks benefactives or recipients, and *for me* more often appears as an evaluative adjunct than does *to me*. Even though *for me* and *to me* are both prepositional phrases containing first person objective pronouns, frequency of their grammatical and pragmatic uses differs considerably because of their distinct constructional and collocational environments. Looking at usage of linguistic expressions in this way allows for closer examination of the functions of grammatical expressions in discourse contexts.

5.2 *For me* as discourse marker?

While *for me* is not as frequent in English conversations as are expressions such as *I mean* and *I think*, it does exhibit some features characteristic of discourse markers. In many of its conversational uses *for me* fulfills pragmatic and interactional functions: it appears as a component of a formulaic expression used to make polite requests (section 7), and it is also used to evaluate and express contrast (section 8). Additionally, *for me* can occur outside the clause structure in initial position as illustrated in (4), which is a central property of discourse markers (Schiffrin 1987; Brinton 1990).

(4) but *for me* it's very difficult to pick up a book about death. (*SBC* 05)

5.3 Classification of *for me* utterances in the database

The three major syntactic types of utterances containing *for me* in the database are displayed in (5).

(5) Syntactic classification of utterances with *for me* (n=395)

a. *FOR* (PREPOSITION) + *me* in predicate final position (adjunct): n=240, 61%
   - and he bought a pair of shoes *for me* (*SWB* 6)
   - it's really really enjoyable *for me*. (*SWB* 319)

b. *FOR* (CONJUNCTION) + *me* (SUBJECT) + *TO* INFINITIVE: n=98, 25%
   - but anyway she has this beautiful plan *for me* to do, (*MICASE* 6)
   - it's hard *for me* to think about food. (*SBC* 24)

c. *FOR* (PREPOSITION) + *me* fronted (disjunct): n=57, 14%
   - sometimes *for me*, they are a whip and a hairshirt. (*SBC* 05)
   - *for me*, that's something I'm not used to. (*SBC* 1)

The most frequent group of utterances in the database is illustrated in (5a); these are expressions in which the prepositional phrase *for me* occurs in predicate-final position, with both lexical and copular verbs. In the group represented in (5b), *for* is traditionally analyzed as a conjunction preceding *me* (the subject of the nonfinite clause) followed by the *to*-infinitive. And in (5c), the prepositional phrase *for me* is independent of the predicate, typically utterance-initial. The more frequent predicate-final group illustrated in (5a) is the focus of the analyses in this paper. These utterances are presented in Table 3, classified by predicate and construction/function.

Table 3. Predicate-final tokens of *for me* by predicate and construction type (n=240)

Type		% Total	Examples
LEXICAL Vs (multiword Vs)	25	10%	*and he's starting to tell me how much he cares for me; but he came back .. looking for me*
LEXICAL Vs (declaratives)	57	24%	*um but he's found some of the* Who *stuff for me; I have the stuff that you did for me last year*
LEXICAL Vs (directives/requests)	23	10%	*give her a hug for me; could you recall those for me*
COPULA + ADJ	68	28%	*it's the starting part that's just so hard for me; so I'm like it'd be easier for me*
COPULA + NP	61	25%	*cause there's a difference for me; so that's a real easy thing for me*
COPULA	6	3%	*it's not for me; I don't know if this is for me*
TOTAL	240	100%	

Table 3 lists three construction types in which *for me* occurs with lexical verbs. The first group includes multiword verbs that contain the particle *for*, such as *care for, look for,* and *wait for*. Included in this category are prefabs with *me*, such as *have feelings for me, works for me,* and *goes for me*. The two other lexical predicate types in Table 3 differ by function: the *declarative* utterances are those in which *me* is typically a benefactive participant, and the *directives/requests* in which *for me* is a component of a formulaic request.

The copular predicates in Table 3 – predicate adjective, predicate nominal constructions, and the formulaic (*it's*) (*not*) *for me* – are more frequent than the lexical predicates, 56 percent and 44 percent, respectively. The rest of the paper provides details of structure and use of the constructions listed in Table 3: declarative utterances in section 6, directive/request forms in section 7, and constructions with copular predicates in section 8.

6. Declarative utterances with benefactive *for me*

6.1 Benefaction

Probably the most studied use of the adjunct *for* + NP (with a human referent) is when it cooccurs with a verb that takes a benefactive argument, and much of this interest has focused on the benefactive alternation in English (Levin 1993), e.g., *Pat baked a cake for Sid* versus *Pat baked Sid a cake*. As discussed by Kittilä (2005) in a recent typological study on beneficiaries and recipients, the expression of benefaction in the world's languages is complex both in meaning and in the coding of case. Heine et al. (1991: 154), however, provide a schematic representation of the benefactive relationship as: X DOES P FOR THE BENEFIT OF Y. This basic construction was used to identify benefactive utterances in the database.

In her study of young children's acquisition of the polysemous prepositions *to* and *for* in British English, Sally Rice (1999: 271) reports that "the very first sense [of *for*] to emerge was benefactive and was always accompanied by the physical transfer of an object". By about 2 years 9 months, the children represented in Rice's data were using *for* for both benefactive of object and of action. Examples of these early uses from Rice (1999: 272) appear in (6) and (7).

(6) BENEFACTIVE OF OBJECT
- *for me*
- it not *for you*
- here are some flowers *for you*

(7) benefactive of action
- Do it *for me*?
- Will you shut that *for me*?

Notice that the expression *for me* appears three times in Rice's presentation, and the other two uses of benefactive *for* occur with the addressee *you*. This illustrates early interactive uses of these forms for young children.

6.2 Verbs with benefactive *for* arguments in the conversational data

Levin's (1993: 48–49) semantic classification of English verbs that take benefactive arguments includes verbs of BUILDING, OBTAINING, CREATING, PREPARING, STEALING, SELECTION, *GET* verbs, and verbs of PERFORMANCE (*dance, paint, recite, sing*). Seventy-four percent of the declarative utterances with lexical verbs (42/57) in the conversations were able to be categorized using Levin's classes

or reasonable extensions of those classes.[3] In these utterances, the subject is an agent who 'obtains', 'creates', or 'gets' something for the first person participant, which functions as a benefactive argument in the clause. Additionally, 66 percent of these verbs are past tense reports. Representative examples from the conversational data appear in (8).

(8) a. BUILD VERBS
– yeah of course he *built* the shelves in the garage *for me* (*SWB* 146)
– well uh why don't you *make* one *for me* and put gold on it, because my parents' fiftieth wedding anniversary is coming up (*SWB* 242)
b. GET VERBS
– um but he's *found* some of the *Who* stuff *for me* (*SWB* 128)
– and he *bought* a pair of shoes *for me* (*SWB* 6)
c. SELECTION VERBS
– how about if I pick for you and you *pick for me* (*SBC* 31)

The presentation of verb types in (8) is somewhat misleading in that it depicts a varied inventory of lexical verb types found in these utterances. In the data, however, 50 percent of the predicates in this group (i.e., those occurring with agent subjects and benefactive *me*) are in fact tokens of one verb lexeme, *do*.[4] Furthermore, 90 percent of these *do* tokens have general or indexical direct objects, e.g., *it, that, all* (*of*) *that, something like that*. Examples of this frequent construction used to express benefaction with a first person singular argument are provided in (9).

(9) AGENT + *DO* + DIRECT OBJECT with general or indexical meaning + *for me*
a. I appreciated her *doing* that *for me*. (*SBC* 47)
b. So he *does* that *for me* and that helps me a lot. (*SWB* 240)
c. No one *did* it *for me*, (*SBC* 44)

Goldberg (1995: 40) notes that *light*, or *general purpose*, verbs (e.g., *go, put, make, do, get*) convey meanings that are similar to argument structure constructions, and that *do* "correspond[s] to the meaning associated with the basic sense of the simple intransitive and/or simple transitive construction". Based on the data reported on here for benefactives occurring with *for me*, not only is it the case that *do* corresponds to the meaning of the *for* benefactive as a (di)transitive construction (X *DOES* P FOR THE BENEFIT OF Y), the lexical item *do* itself occurs as an element in the construction.

3. The 15 tokens of *for me* with lexical predicates not analyzed here do not express typical benefaction, e.g., utterances lacking an agent as in, *I fully expect that that any test results that come back for me would be negative.*

4. *Do* is not classified as a benefactive verb in Levin (1993).

In summary, then, when *for me* occurs in predicate-final position with lexical verbs in declarative constructions in these conversations, it tends to occur in the conventionalized context of AGENT + *DO* PAST + DIRECT OBJECT (with general or indexical meaning). Therefore in these constructions, what is highlighted is the fact that something was (or wasn't) done for the speaker, and not the propositional details of the event (represented by a verb) or of the result of it (the patient-object).

7. Directives and requests with benefactive *for me*

7.1 *For me* as a component of a formulaic speech act

Table 3 shows that 23 utterances in the database contain the prepositional phrase *for me* as a component of a formulaic construction used to make polite requests, e.g., *hey would you mind turning off the lights for me*. Because these uses are associated with service encounters, they are found more frequently in *MICASE* and *AMEX* than in the *Santa Barbara* and *Switchboard* corpora. This use of *for me* appears to be acquired relatively early by children learning (British) English (Rice 1999; see (7)), and it is also part of a formula taught to second language learners. Nattinger & DeCarrico (1992: 52) and Overstreet & Yule (2002: 789) present this use of *for me* as part of a formulaic construction used to make indirect requests in English, illustrated in (10).

(10) MODAL + *YOU* + VP + *FOR ME*
Could you pull up a reservation *for me* (*AMEX* 18)
I want you to grate this cheese *for me* (*SBC* 58)

As discussed by Coulmas (1979), it is common for routine formulae to be tied to ritualized situations of use – in this case, situations of requesting or commanding. And while it is possible that these formulaic uses of *for me* are semantically routinized in the sense that *me* might not be interpreted as explicitly referring to the speaker, a reading of *for me* as a benefactive is still available, though perhaps as a ritualized beneficiary.

7.2 *For me* directives and requests in the conversational data

The directives and requests have different formal properties than do the declarative forms discussed in section 6. Recall that the declaratives tend to be past tense reports containing the light verb *do* or verb types traditionally classified as benefactive. In contrast, all of the verbs in the speech act constructions are present tense, which is not surprising given that these uses are tightly linked to communicative

proceedings. Furthermore, with respect to their predicates, only 26 percent (6/23) of the directive/request constructions contain benefactive verbs. Instead, the majority of these utterances have predicates whose meanings are associated with ongoing activities in the discourse context. For example, six utterances in this group contain verbs related to speaking and communication, e.g., *repeat, spell, recall* (verbally), as illustrated in (11).

(11) DIRECTIVE/REQUEST TOKENS OF *FOR ME* WITH VERBS OF COMMUNICATION
   a. could you uh *recall* those *for me* please (*MICASE* 1-LAB)
   b. maybe it has to do with perceptions Amy.. *put it together for me* Amy (*MICASE* 1-DIS)

Three predicates are verbs of movement or body stance, as shown in (12), and the eight remaining uses are requests and directives related to activities in the context. These are illustrated in (13).

(12) DIRECTIVE/REQUEST TOKENS OF *FOR ME* WITH VERBS OF MOVEMENT
   a. *bend over for me* Darren, (*SBC* 57) [judo class demonstration]
   b. we don't you know say *get down* and *do* a hundred push ups *for me* now you little guy (*SWB* 189) [hypothetical direct quote]

(13) DIRECTIVE/REQUEST TOKENS OF *FOR ME* WITH VERBS LINKED TO CURRENT ACTIVITIES
   a. can someone *run down* the axiom of existence *for me* real quick (*MICASE* 1-SGR)
   b. could you *pull up* a reservation *for me* please (*AMEX*)

While verbs of speaking and body stance or movement are not typical predicates occurring with benefactive arguments, an interpretation of the speaker as beneficiary in these utterances is still available. Because these predicates refer to activities in the discourse context, the speaker does 'benefit' from the addressee's compliance with her request – not, however, as a propositional participant represented as an adjunct in a clause – but as an on-site speech act participant.

For example, speaker S3 in (14) works at a university media desk checking out equipment. Using the indirect request construction with *for me* in line 4, *okay I just need you to fill out the bottom part for me please*, she instructs the patron to complete the transaction by signing a form.

(14) 1. S24: can you check this out too
   2. S3: sure
   3. can you come up to this computer
   4. → okay I just need you to fill out the bottom part *for me* please
   5. and I'll need to see your M-card (*MICASE* 1-SVC)

The social role of the speaker, S3, in this interaction is that of a participant in a service encounter, and in order for her to perform her job, the patron, participant S24, must sign the form. The benefit to S3, however, is not in the specific details of the signing itself, in the same way, for example, that a beneficiary of an act of 'buying' or 'building' receives a tangible result of that action (a case of *concrete benefaction*, Kittilä 2005: 273). Nor is the benefit to Speaker S3 *substitutive* (Kittilä 2005: 273) as it would be if she were the one in need of the equipment but happened to be holding several packages and was unable to sign the form for herself. Instead, in this case, the benefit to the speaker issuing the request using *for me* is in the conventional necessity that the patron must sign the form as part of the socially situated process of checking out AV equipment at this university media desk. The speaker benefits, certainly, but it is a ritualized benefit related to her social role in a routine encounter.

These tokens of directives and requests with *for me* are examples of Coulmas's (1979) *ritual formulae*, expressions whose uses are closely tied to recurrent social situations. And the analyses presented in this section suggest that the formal properties of these conversational utterances – present tense, lexical verbs with meanings that refer to activities ongoing in the discourse context – are consistent with their functions in discourse.

8. Copular predicates with evaluative *for me*

8.1 Formal properties of copular constructions with *for me* in the conversational data

Recall from Table 3 that 56 percent of the predicate-final tokens of *for me* occur with copular predicates, primarily predicate adjective and predicate nominal constructions. Other studies have also found that copular clauses are highly frequent in English conversations (Thompson & Hopper 2001; Scheibman 2002). Examples of these copular constructions with *for me* are provided in (15).

(15) Copular constructions with *for me* (n = 135)
   a. Predicate adjective constructions(n=68)
   – 3RD PERSON INDEXICAL SUBJ + *BE* + ADJP (often *hard + for me)*
   – yes it's uh it is sure tough *for me* (*SWB* 299)
   – so I'm like it'd be easier *for me* (*MICASE* 2-SGR)
   b. Predicate nominal constructions (n=61)
   – 3RD PERSON INDEXICAL SUBJ + *BE* + NP with general meaning + *for me*
   – cause this is really a lot of work *for me* (*MICASE* 3-TOU)
   – so it's not a problem *for me* (*SWB* 372)

c. 3RD PERSON INDEXICAL SUBJ + *BE* (*NOT/N'T*) + *for me* (n=6)
– it's not .. *for me.* (*SBC* 04)
– whoops you know it's not *for me* (*SWB* 392)
– I don't know if this is *for me,* (*SBC* 44)

Copular utterances with *for me* adjuncts in the data are constructions whose uses are characterized by: (1) the presence of third person singular subjects (100 percent of the tokens), which are overwhelmingly inanimate (95 percent) and pronominal (80 percent), primarily tokens of *it*, *that*, and *this*, with (2) frequent occurrence of present tense predicates (75 percent of the tokens). With respect to the predicate adjective constructions, 31 percent (21/68) of the lexical adjectives occurring in these expressions are tokens of *hard* or adjectives with related meanings (e.g., *tough, difficult, easy*). Additionally, the head nouns in the predicate nominal clauses of utterances illustrated in (15b) tend to be lexical items with general meanings, such as *experience, work, time, thing* – words which provide open semantic slots for participants to situate and evaluate experiences in discourse. As is the case for the declarative and request/directive utterances, these routine copular expressions containing *for me* include grammatical, lexical, and semantic elements as part of their constructional material (Goldberg 2003; Bybee & Eddington 2006).

Unlike the declarative and directive utterances with lexical verbs discussed in sections 6 and 7, however, the first person pronoun *me* in the copular constructions is not benefactive. Not only do copular verbs not designate processes of '*doing* P for the benefit of Y', but 95 percent of these utterances have subjects with inanimate referents – entities which could not be semantic agents. In these frequent utterances, then, *for me* marks an evaluating speaker, not a beneficiary.

8.2 Evaluative properties

To broaden examination of the evaluative character of these copular constructions, formal evaluative features – specifically markers of comparison and intensity – were identified for each utterance (Labov 1972; Thompson & Hunston 2000). Examples appear in (16).

(16) Examples of markers of comparison and intensity in copular clauses with *for me*
– negatives: ***not*** *practical,* ***not*** *as comfortable*
– intensifiers and other comparators: ***too*** *sober, easier, short* ***enough, really, real, very***

Eighty-one percent of the predicate adjective expressions (55/68) and 74 percent of the predicate nominal constructions (45/61) contain a least one comparator or intensifier. Indeed, only 29 tokens from the two main copular groups (22 percent) do not contain these evaluative features.[5]

To compare these copular predicates with *for me* with copular clauses without *for me*, I conducted a cursory examination of evaluative features of copular utterances without *for me* from another conversational database (Scheibman 2002). In these data, only 38 percent (57/150) of the predicate adjective clauses (compared to 81 percent of *for me* tokens, and 15 percent (31/210) of the predicate nominal clauses (compared to 74 percent of *for me* tokens) contain at least one of the evaluative features. This difference suggests that the copular constructions with *for me* formally reflect their evaluative functions in interactive discourse.

8.3 Evaluative and metapragmatic functions in conversations

The basic construction characterizing the conversational usages of copular utterances with *for me*, then, consists of an inanimate third person pronominal subject with a present tense form of the copula (typically *to be*). For the adjectival predications, there is a tendency for the lexical adjectives to be tokens of the word *hard* or items with related meanings, and for the nominal predications, the head noun of the subject complement is semantically general. Additionally as a group, 78 percent of these utterances include markers of comparison and intensity (evaluative features). Based on expectations in usage-based linguistics and work by Coulmas and others who propose that routinized linguistic expressions are associated with routine social uses, this section looks at the conventionalized functions that are performed by these constructions in the conversations.

In an analysis of evaluative *to*-adjuncts and *for*-adjuncts in English, Kudrnáčová (1987: 131) writes that "[a]n important role in an evaluative predication is played by the evaluator, the bearer of the evaluative attitude or reaction". In the case of the evaluative adjunct *for me*, the evaluator is also the speaker which gives the evaluation pragmatic functions in interaction.[6] In her comparative analysis of *to*- and *for*-adjuncts, Kudrnáčová suggests that *for*-adjuncts refer to *external conditions* in the evaluator's mind, while *to*-adjuncts refer to *internal conditions*. To illustrate this contrast, Kudrnáčová offers a pair of adjectival predications with infinitive clauses reproduced in (17).

5. Some of the predicate adjective expressions that did not contain evaluative features did include adjectives with lexically expressed intensity and comparison, e.g., *huge* as in this example where the speaker was discussing having lost $100 gambling: *but you know for us that's huge for me but you know some people do it every weekend* (SWB 137).

6. Kudrnáčová doesn't explicitly comment on the role of first person in these expressions; however, the majority of her examples are *to me* and *for me*.

(17) Kudrnáčová (1987: 132)
It is important *for me* to read technical books.
It is important *to me* to read technical books.

Though Kudrnáčová (1987) does not analyze this particular example, the *for* version lends itself to an interpretation that the importance of reading the books has to do with some external condition or requirement, such as one required for an educational course or licensure program. In contrast, the *to* utterance suggests that reading the books is perhaps important to the speaker's intellectual life in some way (internal conditions).

Given this idea that *for*-adjuncts "refer to external circumstances in which the evaluator finds himself" (Kudrnáčová 1987: 132), *for me* adjuncts in these copular constructions might be said to be used by participants to evaluate their experiences and reactions based on expectations related to social and cultural conventions (i.e., external conditions). However, because these usages occur in the dialogic and collaborative contexts of interactive discourse, participants use these evaluative constructions with *for me* to assess a reaction or experience relative to – not only social and cultural expectations in a global sense – but to expectations assumed to be shared by the other interactants. These evaluative constructions with first person elements, then, have *metapragmatic* functions in discourse in the sense that are used by speakers to convey awareness that their evaluations may be in contrast to, or at least relative to, those of other discourse participants.

Verschueren (1995: 367) defines *metapragmatics* as "[t]he systematic study of indicators of the language user's reflexive awareness of what is involved in a usage event, i.e., the study of a metalevel at which verbal communication is self-referential to various degrees". The presence of predicate-final *for me* in these evaluative utterances shows a reflexive awareness of both the speaker's stance and assumed norms informing that stance. In conversations, interactants tend to use these constructions to situate themselves favorably as they convey assessments and reactions that they view as contrasting with social conventions (e.g., cultural notions of fashion, modes of relating or earning a living, competence in one's job, or how best to conduct commercial transactions). These evaluative comparisons can be low stakes for a speaker, e.g., that he likes his orange juice particularly cold (18), or that he finds wearing sweatpants more comfortable than wearing a tie (19).

(18) I keep it really cool.
so, it's cold enough *for me*. (*SBC* 28)

(19) I like to wear like sweat pants because it's more comfortable *for me*. (*SWB* 23)

These metapragmatic presentations, however, can also be interactionally consequential. In the conversational excerpt in (20), the speaker, Sharon, is a beginning elementary school teacher who has so far not had an easy time in her new job. Immediately before the excerpt in (20), Sharon told the other participants that among other challenges at work, she was given a class composed of both third graders and fourth graders – a situation that requires more experienced teachers to manage handily (line 5). Her evaluation in line 8, *it's like really hard for me*, situates her stance toward her struggles at school for the others participants.

(20) (*Raging bureaucracy SBC*0004 572.79–581.61)

1.	SHARON:	it's,
2.		well,
3.		[it's something],
4.	KATHY:	[It's twice] as [much work for you].
5.	SHARON:	[it's something for] experienced teachers.
6.		It's not .. for me.
7.		.. You know,
8.	→	it's like really hard *for me*.
9.		Because,
10.		you know,
11.		[and then –
12.	KATHY:	[You have to –
13.	SHARON:	and then],
14.	KATHY:	That's what I was] doing when I was student teaching,
15.	SHARON:	Well they didn't even give me any texts
16.		I mean I was the only teacher,
17.		in the whole school,
18.		who did not have textbooks.

In line 14, another participant, Kathy, perhaps to show solidarity, responds by saying that she also taught a split class when she was student teaching. However, in offering this information, Kathy ends up matching her experience with Sharon's, which has the effect of normalizing (perhaps, neutralizing) the challenges of the external situation (teaching split classes) that Sharon has situated as a reference point in her own presentation of her struggles to the group. Subsequently in lines 15–18, Sharon augments her position by noting that she was also the only teacher in the school who did not have textbooks. In this extended example, the speaker's attempt to favorably situate her challenges with external circumstances for other interactants was not without interactional complexity.

The frequent occurrence of the prefab (*it's*) *hard for me* suggests that it is not unusual for participants in these conversations to cast their evaluations of their experiences as not meeting expectations in particular ways. But interactants also present their experiences as positive relative to assumed norms. In the short narrative in (21), the speaker, Pamela, states in lines 15 and 16 that her terrible first

marriage, which might be assumed to be a devastating experience, instead affected Pamela positively because in the end it motivated her to become more independent than she was when she was in the marriage. The speaker uses *for me* in this example to assert and convey to the other participant her awareness that her individual experience is in contrast to conventional expectation, and in this case, it is positively evaluated.

(21) (*A book about death SBC*0005 45.88–105.68)[7]

1.	PAMELA:		the fact of the matter is,
2.			that the marriage itself,
3.			I mean as hellish as it was,
4.			.. it's like it pulled me under,
5.			like a giant octopus,
6.			or a giant,
7.			giant shark.
8.			And it pulled me all the way under.
9.			And then,
10.			and there I was,
11.			it was like the silent scream,
12.			and then,
13.			then I found that .. I was on my own two feet again.
14.			And it really was –
15.		→	s- what was hell in that .. that marriage became,
16.		→	became a way out *for me*.

In interactive contexts, then, these copular constructions with *for me* are used by participants to position their evaluations of their own experiences relative to social and cultural expectations. The formal properties of these utterances are consistent with these uses, such as the frequency of evaluative features and lexical items that mark challenge (e.g., *hard*) or those that designate general situations (e.g., *experience, problem, way out*). The first person element in these constructions indexes the speaker's awareness of normative expectations and contributes to these expressions' frequent pragmatic functions in conversations.

9. Summaries and conclusions

The analyses offered in this paper suggest that the expression *for me* is entrenched in multiple sites in discourse, each site characterized by its own constructions

7. The copula *to become* appears in this example.

and uses. In its three most frequent discourse contexts, predicate-final *for me* occurs as part of constructions which contain grammatical, lexical, and (often, general) semantic material. The prepositional phrase is a benefactive adjunct in declarative utterances with lexical predicates (primarily the verb *do*), a formulaic component of polite directives and requests which indexes the speaker as a ritualized beneficiary, and a marker of evaluation in copular clauses which perform metapragmatic functions. What these analyses suggest is that not only is a given expression conventionalized, but its linguistic contexts and its interactional uses are also conventionalized.

The presence of the first person pronoun *me* largely contributes to *for me's* frequent pragmatic functions. In the majority of its uses, *for me* is not propositionally robust. For example, *for me* does not add much content to the indirect request *I was wondering if you could handle some reservations for me*.[8] Similarly, in the copular constructions, addressees are likely to assume that when a speaker produces an utterance such as, *It was very relaxing for me*, that the assertion represents the speaker's own evaluation, even if it hadn't been marked as such with *for me*. And while it is the case that *for me* in the declarative utterances with lexical verbs adds information to the clause by introducing a benefactive argument, due to the presence of *do* and direct objects with general meanings in these constructions, what ends up being highlighted in these utterances is the fact that something was or wasn't done for the speaker, and not the details of the propositional event. In these data, then, *for me* frequently fulfills pragmatic functions. This is in line with previous studies that have shown that expressions with first person elements are particularly open to performing reflexive and interactive tasks, uses which contribute to their frequency and routinization in conversational discourse.

As discussed in the introductory portion of this paper, studies in usage-based linguistics and formulaic language research have suggested that routine expressions are linked to social and communicative situations. While it is the case that discourse is replete with conventionalized structures and formulas, participants' actual uses of these forms are sensitive to details of context and interaction (e.g., prior discourse, characteristics of the situation and participants, shared knowledge and expectations). Such variation creates analytical challenges both for linking formal linguistic properties to recurring communicative and social functions, and for generalizing functions across contexts. From the perspective of usage-based theories, however, it is these repeated cooccurrences of interaction and linguistic form that 'produce' the patterns and regularities documented by analysts in their work.

8. In some directives/requests, however, in which *for me* marks *substitutive benefaction*, e.g., *give her a hug for me*, mention of the speaker is more informative.

In his discussion of emergence of language structure in discourse, Hopper (1998: 156) characterizes grammar (a model of formulaic language) in this way:

> Its forms are not fixed templates but emerge out of face-to-face interaction in ways that reflect the individual speakers' past experience of these forms, and their assessment of the present context, including especially their interlocutors, whose experiences and assessments may be quite different.

This study of one routinized phrase, *for me*, takes seriously Hopper's proposal that linguistic structure is continually "becoming", contingent on communicative and interactive exigencies. Indeed, to have simply treated *for me* as a rank-and-file member of the class of English prepositional phrases would not have put in focus the kinds of distributional and functional details that result from observing an expression's multiple uses by participants, as reported on here. In particular, the examination of *for me* draws attention to the indexical character of utterance types and their formal properties found in discourse, and highlights ways in which language structure and recurring social actions can be viewed as mutually dependent.

Corpora

AMEX corpus of human-human air travel planning dialogs. 1989. SRI International.

Du Bois, J.W., Chafe, W., Meyer, C. and Thompson, S.A. 2000. *Santa Barbara Corpus of Spoken American English, Part I*. 3 CD-ROMs. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Du Bois, J.W., Chafe, W., Meyer, C. and Thompson, S.A 2003. *Santa Barbara Corpus of Spoken American English*, Part II. 1 DVD. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Du Bois, J.W. and Englebretson, R. 2004. *Santa Barbara Corpus of Spoken American English*, Part III. 1 DVD. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Du Bois, J.W. and Englebretson, R. 2005. *Santa Barbara Corpus of Spoken American English*, Part IV. 1 DVD. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Godfrey, J.J. and Holliman, E. 1993. *Switchboard-1 Transcripts*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Godfrey, J.J. and Holliman, E. 1997. *Switchboard-1 Release 2*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Simpson, R.C., Briggs, S.L. Ovens, J. and Swales, J.M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

References

Aaron, Jessi Alana & Rena Torres Cacoullos. 2005. Quantitative measures of subjectification: A variationist study of Spanish *salir(se)*. *Cognitive Linguistics* 16: 607–633.

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

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

Brinton, Laurel J. 1990. The development of discourse markers in English. *Historical linguistics and philology*, Jacek Fisiak (ed.), 45–71. Berlin: Mouton de Gruyter.

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

Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82: 323–355.

Bybee, Joan & Paul J. Hopper. (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37: 575–596.

Coulmas, Florian. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3: 239–266.

Coulmas, Florian. 1981. Introduction: Conversational routine. *Conversational routine: Explorations in standardized communication situations and prepatterned speech*, F. Coulmas (ed.), 1–17. The Hague: Mouton.

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7: 37–77.

Du Bois, John, Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino. 1993. Outline of discourse transcription. *Talking data: Transcription and coding in discourse research*, Jane A. Edwards & Martin D. Lampert (eds.), 45–89. Hillsdale NJ: Lawrence Erlbaum Associates.

Fillmore, Charles J. 1977. The need for a frame semantics within linguistics. *Statistical methods in linguistics*, H. Kalgren (ed.), 5–29. Stockholm: Scriptor.

Fox, Barbara & Sandra A. Thompson. 2007. Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language* 31: 293–326.

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

Heine, Bernd, Ulrike Claudi & Friederike Hünnemeyer. 1991. *Grammaticalization: A conceptual framework*. Chicago IL: The University of Chicago Press.

Hopper, Paul J. 1987. Emergent grammar. *Berkeley Linguistics Society* 13: 139–157.

Hopper, Paul J. 1997. When 'grammar' and discourse clash: The problem of source conflicts. *Essays on language function and language type*, J. Bybee, J. Haiman & S.A. Thompson (eds.), 231–247. Amsterdam: John Benjamins.

Hopper, Paul J. 1998. Emergent grammar. *The new psychology of language: Cognitive and functional approaches to language structure*, M. Tomasello (ed.), 154–175. Mahwah NJ: Lawrence Erlbaum Associates.

Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation: A description of its interactional functions, with a focus on* I think. Amsterdam: John Benjamins.

Kittilä, Seppo. 2005. Recipient-prominence vs. beneficiary-prominence. *Linguistic Typology* 9: 269–97.

Kudrnáčová, Naděžda. 1987. A note on the *to*-adjunct and the *for*-adjunct on their evaluative use. *Brno Studies in English* 17: 131–140.

Labov, William 1972. *Language in the inner city: Studies in the Black English vernacular*. Philadelphia PA: University of Pennsylvania.

Levin, Beth. 1993. English verb classes and alternations: A preliminary investigation. Chicago IL: The University of Chicago Press.

Nattinger, James R. & Jeanette S DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: OUP.

Ono, Tsuyoshi & Sandra A. Thompson. 2003. Japanese (w)atashi/ore/boku 'I': They're not just pronouns. *Cognitive Linguistics* 14: 321–347.

Overstreet, Maryann & George Yule. 2002. The metapragmatics of *and everything*. *Journal of Pragmatics* 34: 785–794

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Native-like selection and nativelike fluency. *Language and communication*, J.C. Richards & R.W. Schmidt (eds.), 191–226. London: Longman.

Rice, Sally. 1999. Patterns of acquisition in the emerging mental lexicon. *Brain and Language* 68: 268–276.

Scheibman, Joanne. 2000. *I dunno…* A usage-based account of the phonological reduction of *don't* in American English conversation. *Journal of Pragmatics* 32: 105–124.

Scheibman, Joanne. 2002. *Point of view and grammar: Structural patterns of subjectivity in American English conversation*. Amsterdam: John Benjamins.

Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: CUP.

Tao, Hongyin. 2001. Discovering the usual with corpora: The case of *remember*. *Corpus linguistics in North America: Selections from the 1999 Symposium*, R. Simpson & J. Swales (eds.), 116–114. Ann Arbor MI: University of Michigan.

Thompson, Geoff & Susan Hunston. 2000. Evaluation: An introduction. *Evaluation in text: Authorial stance and the construction of discourse*, S. Hunston & G. Thompson (eds.), 1–27. Oxford: OUP.

Thompson, Sandra A. 2002. Object complements and conversation: Towards a realistic account. *Studies in Language* 26: 125–164.

Thompson, Sandra A. & Elizabeth Couper-Kuhlen. 2005. The clause as a locus of grammar and interaction. *Discourse Studies* 7: 481–506.

Thompson, Sandra A. & Paul J. Hopper. 2001. Transitivity, clause structure, and argument structure: Evidence from conversation. *Frequency and emergence of linguistic structure*, J. Bybee & P. Hopper (eds.), 28–60. Amsterdam: John Benjamins.

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

Tsui, Amy B.M. 1991. The pragmatic functions of *I don't know*. *Text* 11: 607–622.

Verschueren, Jef. 1995 Metapragmatics. *Handbook of pragmatics manual*, J. Verschueren, J.-O. Östman & J. Blommaert (eds.), 367–371. Amsterdam: John Benjamins.

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