The idiom principle and the open choice principle

BRITT ERMAN and BEATRICE WARREN

Abstract

The assumptions forming the basis of this study are that the language user has available a number of more-or-less preconstructed phrases and that the production of texts involves alternation between word-for-word combinations—which we refer to as adherence to the open choice principle (after Sinclair 1991)—and preconstructed multi-word combinations, which we refer to as making use of the idiom principle (again after Sinclair).

The main aim of the study is to gain an impression of the impact that this alternation has on the structure of texts. Therefore a mode of analysis has been worked out revealing how multi-word combinations combine with each other and with words combined according to the open choice principle. This is the main contribution of the study. Another important contribution is the revelation that there is a large amount of prefabricated language in both spoken and written texts (on average around half of the texts), which makes it impossible to consider idioms and other multi-word combinations as marginal phenomena.

Keywords: prefabs; idiomaticity; compositionality vs. non-compositionality; word-for-word production vs. storage; structure of texts.

1. Introduction

1.1. Theoretical background

The traditional view that production of utterances involves the organization of stored primitives in terms of a relatively large number of rules was attacked at least as early as 1974 by Bolinger in a lecture.1 ‘Speakers do at least as much remembering as they do putting together’ (1976: 2), he claimed, and suggested that, in view of the fact that the human brain is capable of extensive memory storage, it would be more natural to work
with the following hypothesis: we store a large number of complex items which we manipulate with comparatively simple operations. Sinclair (1991: 110) expresses the same idea in the following way:

the principle of idiom is that a language user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.

He contrasts the principle of idiom with the open choice principle, which then represents the traditional assumption that practically each position in a clause offers a choice. Another attack on the traditional approach comes from Pawley and Syder (1983), who point out that it cannot account for nativelike selection (idiomaticity) or fluency. More recently, Fillmore, Kay, and O'Connor (1988) and Jackendoff (1997) have suggested theories in which allowances are made for idiomaticity in language; we will return to these in section 5.

Observations of this kind have inspired an extensive number of studies concerned with different aspects of ‘chunking’. Conventionalized phrases are naturally of particular interest to lexicographers (e.g., Kjellmer 1994; Cowie et al. 1975–1983; Benson et al. 1993) and to linguists interested in first and second language acquisition (e.g., Wong-Fillmore 1976; Peters 1983; Granger 1996; Bolander 1989; Weinert 1995), but also to linguists working with machine translations (Ahrenberg and Merkel, to appear; Sigurd et al. 1992), with text-to-speech conversion (Lindberg 1996), conversational routines (e.g., Aijmer 1996; Coulmas 1981), or recurrent strings or collocations (Altenberg 1996a, 1996b; Kjellmer 1994; Sinclair and Renouf 1991). The main concern of the present study is the impact that prefabricated language has on the production and structure of the text, whether spoken or written.

1.2. Aims of the study

The basic assumption of the project is that in producing utterances the language user alternates between the open choice principle and the idiom principle. Traditional structural analysis displays how words are combined to make grammatical phrases but ignores the fact that words also combine to make multi-word composites (‘prefabs’), which are not necessarily convergent with grammatical phrases. Therefore, a mode of analysis has been developed here with the objective of making the alternation between prefabricated and non-prefabricated combinations manifest. The purpose of this method of analysis is, then, to give an impression of the impact that prefabricated language has on the structure of a text and on the effort involved in encoding and decoding it.
An additional aim is to give a more realistic representation of the mental lexicon.

More specifically, the questions the analysis addresses are:

i. What is the average proportion of prefabs in texts; and in a text of X number of words, how many single choices will have been made on average?

ii. Does the frequency of prefabs depend on whether the text is spoken or written?

iii. How do prefabs combine with each other and with words combined according to the open choice principle?

iv. What is the distribution of lengths (average and maximum, etc.) of prefabs and non-prefabricated strings?

v. Do prefabs serve certain main functions such as lexical, pragmatic, and grammatical ones? Presumably it is possible to relate certain types of prefabs to certain types of text. Other prefabs must be stylistically neutral. Which and why?

vi. How and why can prefabs be varied?

1.3. Plan of the article

In the following our first concern will be establishing definitions of and criteria for prefabs, which is the topic of section 2. In section 3, we give an account of the material we studied and of the mode of analysis that we have developed. Section 4 presents the results of the study. Apart from statistics, this section contains a discussion of different types of prefabs, of how they can be combined, and of frequencies of retrievals in a text. We conclude by considering practical and theoretical implications of the study and by suggesting a model for the storage and production of language which takes into account all lexical knowledge, including multi-word units.

2. Prefabs

2.1. Definitions and criteria

The definition of prefab adhered to in this study is as follows:

A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization.

Definitions along these lines can be found in a number of papers concerned with phraseology (for an overview, see Weinert 1995). Nattinger
Britt Erman and Beatrice Warren and DeCarrico (1993: 558–567; our emphasis) offer a particularly informative version:

... lexical phrases are chunks of language of varying length, conventionalized structures that occur more frequently and have more idiomatically determined meaning than language that is put together each time.

Our definition stipulates that a prefab must be a combination of words. For practical reasons, we decided that a word is whatever is represented orthographically as a word. Therefore, teacup spelt as one word would not be considered a prefab, but tea cup spelt as two words would. The definition further stipulates that the combination must manifest some feature of conventionalization, which would be evidence—although no proof—that it is memorized. This in turn suggested the criterion most frequently made use of in the present study to ascertain 'prefab status', namely restricted exchangeability. By restricted exchangeability is meant that at least one member of the prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity. For instance, good friends in they are good friends cannot be changed into nice friends without losing the implication of reciprocity; not bad (meaning 'good') cannot be changed into *not lousy without a change of meaning and loss of idiomaticity; I can't see a thing cannot be *I can't see an object without loss of the non-literal hyperbolic meaning; I'm afraid—a pragmatic prefab used to soften a piece of bad news cannot be *I'm scared or frightened.

Restricted exchangeability may also imply the blocking of certain syntactic variability which is normally possible: I guess—a pragmatic epistemological prefab—cannot be negated (*I don't guess); It will do cannot lose its auxiliary (*It does); up here cannot have a reversed order (*here up).

It should be pointed out that this stipulation—i.e., that for anything to be a prefab the choice of one word must determine or, at least definitely restrict, the choice of at least one other, normally adjacent, word—excludes from consideration constructions which are no doubt memorized such as make somebody do something, but which contains only one lexically specified item (somebody, do, and something all represent unrestricted choices of words). In other words, prefabs are not the same thing as constructions in Construction Grammar, since constructions can be non-lexically specified and can consist of one word. (See Kay 1997, in particular pp. 123–131). (In this connection we may add that the aims of the present study are more modest than those of Construction Grammar. We do not have the ambition to describe the phonological, syntactic, semantic, and pragmatic characteristics of the individual prefabs.)
This stipulation also excludes from consideration completely compositional collocations such as *dark night*, in which neither *dark* determines the choice of *night*, nor *night* the choice of *dark*. Compare in this respect the combination *choose words*, which, although it is transparently motivated, we consider a prefab since, provided we wish to refer to the mental activity involved in producing utterances, the expression cannot be *pick words, select words, excerpt words* without being marked or unidiomatic. *Choose words* is an example of a prefab which is syntactically fully compositional, but not semantically so.²

However, as has already been admitted, even if the criterion of restricted exchangeability is satisfied, there is no absolute proof that a particular combination of orthographic words is ready-made. It is quite possible that, say, *I am afraid* implying 'I regret to have to inform you' was composed word for word by a speaker who simply happened to make the idiomatically correct choice of words.

The application of the criterion of restricted exchangeability can sometimes fail in another respect. It is not always possible to find acceptable synonyms for the members of a candidate for 'prefab status'. Consider, for instance, *very well* in *I managed very well*, which we could not test since we failed to think of synonyms close enough in meaning. *Very well* was nevertheless accepted as a prefab, as were some other combinations for which we have no other criterion than our intuition.

The identification of prefabs is difficult. There are two main reasons why this is so. One is that what is a prefab to some members of a language community need not be a prefab to all members. Some prefabs will be known to practically all native and fluent speakers; others will be more limited in dispersion and entrenchment. This is natural in view of the fact that conventionalization is a gradual process and that the usefulness of prefabs may vary. Prefabs, it must be remembered, are not like phonemes and morphemes, or noun phrases and verb phrases. They are probabilistic, some more than others. The other reason that prefabs are difficult to identify is that they can easily be overlooked. Some of them appear at first sight to be completely transparent combinations of words which reveal themselves to be idiomatic or noncompositional (in a strict sense) only on closer examination. Fortunately, however, not all prefabs are inconspicuous. Idioms, compounds (provided they are spelt as two words and are well established), habitual collocations (provided they are noncompositional) and prepositional and phrasal verbs are all fairly easily identifiable and indisputable examples of prefabs. Nevertheless, one of the findings of this study is that the identification of 'all and only' the prefabs in a text is in practice impossible.
In view of these uncertainties, we do not claim that our analysis is the only possible one, or that our statistical results are precise (they should rather be looked upon as approximations). Nevertheless we claim that our results give a fairly reliable impression (and an interesting one) of the impact that prefabricated language has on the structure of a text. The fact that there is some indeterminacy in our analysis reflects that there is some indeterminacy among language users as to what would be seen as common and acceptable combinations and what would not.

We are of course aware that manifestability is a virtue in any investigation of a scientific nature and that the greater the degree of manifestability, the more reliable the study will be. However, rigid insistence on precise manifestability must not exclude from study probabilistic phenomena in language, phenomena which can be of the utmost relevance if we are to account for the production, interpretation, and acquisition of language.

3. Mode of analysis

The material we analyzed consists of nineteen extracts: seven of 600 to 800 words from different texts in The London Lund Corpus of Spoken English (LLC) and ten extracts of 100 to 400 words from different text types in the Lancaster-Oslo-Bergen corpus (LOB), representing written English, plus two 400-word extracts from two versions of Goldilocks.

Each word in a text is thought to fill a slot. A slot can be filled by a word chosen according to the open choice principle or a word which is part of a prefab. In our analysis, words which are not parts of prefabs are removed and replaced by dashes (— for each word); words which are part of prefabs are left. This is meant to give a visual impression of the alternation between prefabs and non-prefabricated elements. The beginning and end of a prefab is marked by a slash:

(1) B /you know// I went to some seminars/ A —— —— /a waste of time//are they/ [LLC 1.5. 13–16]

(Here the notation in brackets following the example indicates the corpus, and the text and tone unit numbers.)

Frequently a prefab has one or several open slots, i.e., slots that must be filled with lexical material for the prefab to be complete, but which can be filled by a practically unlimited number of words. Such words are
The idiom and open choice principles

not removed but are italicized and reduced in size, as in the second prefab in example (1) and in the third in example (2):

(2) B — — — — — —/or not/  
    A — — — — — /but only// from Kyd onwards/  
    [LLC 1.5. 17–21]

These words are ignored when the number of slots filled by parts of prefabs is counted. Not infrequently, open slots are filled by prefabs. To indicate this, we make use of square brackets and do not reduce the size of the prefab:

(3) /To some extent/ /[the answer to this question] depends on [one's answer to question 1 (above)]/  
    [J34 013–021]

(Here, J34 ‘013–021’ identifies the context of the example.)

We see in this example and in (1) that a prefab may have a variable member: go to seminars/lectures/classes; to some/a certain extent; the/an etc. answer to this/that etc. question. If the variability is judged to be restricted as in these examples, it is ignored in the prefab analysis of the text, but not in estimating number of choices, as we will see later. Nor is it normally ignored in the dictionary, i.e., in the list of prefabs compiled after a text has been analyzed, where it is indicated in the following way: break a rule/law etc.; go to a seminar/lecture etc. (However, certain types of predictable variation such as that between definite and indefinite articles are also ignored in the dictionary.) Provided it is possible, open slots are represented in the dictionary by sth or sb (something or somebody), as in sth depends on sth.

Prefabs are often extendible:

(4) (quite) all right  
    due (mainly) to sth

Extensions, which are non-obligatory, are put within parenthesis. Some may be common (although optional) members of the prefab in question (as in [4]); others are not considered part of the prefab. These are reduced in size and disregarded.

Extensions (in fact in example [5]) as well as obligatory parts of prefabs (a few and that sort of) may of course be prefabs themselves:

(5) a. /What ([in fact]) did you do?/  
    b. /be ([a few] minutes) late/  
    c. /lead [that sort of] life/
So far only lexical prefabs have been discussed. There are, however, also function prefabs. These serve either grammatical or pragmatic functions. In the dictionary the different types of prefabs are listed separately and the particular functions of the grammatical and the pragmatic prefabs are indicated. Typical grammatical function labels are quantifiers (*a few, a great deal of, etc.*), links (*for instance, instead of, either ... or*), determiners (*some of, that sort of*), proforms (*and so on*), tense (*be going to*), modals (*be supposed to, have got to*), introductors (*there is|are*), and aspect (*used to*). Typical pragmatic function labels include discourse markers (*and then*), feedback signals (*yeah quite*), performative routines (*thank you, good evening*), hedges (*I should think*). In most cases, deciding which type a prefab represents is relatively unproblematic, but there are naturally also a number of debatable cases, as will become evident.

The existence of lexical, pragmatic, and grammatical prefabs reflect different reasons for the conventionalization of phrases. According to Pawley and Syder (1983), it is natural that there are standard ways of referring to standard situations and phenomena in a culture. Phrases of that kind become lexical expressions. Similarly, there will be standard ways of expressing oneself in standard social interaction and standard ways of monitoring texts, giving rise to pragmatic and grammatical expressions. Therefore postulating the existence of lexical, grammatical, and pragmatic prefabs can hardly be considered controversial. The postulation of a fourth type of prefab, viz. reducibles, may be more debatable. Examples of reducibles include *it's, I'm, you're, that's, hasn't let's*, etc. Reducibles have no obvious functional rationale and normally no idiomatic meaning. Nevertheless, it is postulated that they have become memorized units because the two members co-occur so frequently that the retrieval of one of them calls up the other. In written English, reducibles will normally be spelt as two words, but we have found it consistent to also consider these full versions as prefabs, unless there are clear indications to the contrary. As already suggested, we will discuss the different prefab types in greater detail in section 4.

Having established the number and type and proportion of prefabs in the text under analysis, the next step is to determine the number of choices made in producing the text. Underlinings are used to mark a choice, i.e., each choice is underlined by a solid line:

(6) To the best of + my + knowledge, there is no record of a society which has used literacy for + the profane and imaginative + purposes and which has + not + produced books dealing with sexual topics;

[23 choices; 33 slots; G77 001–010]
Plus signs as in ‘to the best of + my + knowledge’ indicate that this string involves two choices (the prefab and the variable determiner) and not three. As can be seen from example (6), the choice of tense has been ignored. That is to say, *produces, produced, has/had produced, will produce* are considered to be one choice irrespective of whether an auxiliary is involved or not. Our reasons for these policies are given in section 4.5.

Finally, the number of lexical choices is determined. In our analyses lexical choices are marked by the use of bold print:

(7) **To the best of + my + knowledge,** there is no **record of a society** which has used literacy for + the **profane** and **imaginative** + purposes and which has + not + **produced books dealing with sexual topics**;

[13 lexical choices; G77 001–010]

Some additional illustrations of this mode of analysis are given in Appendix 2.

4. Results

4.1. Statistics

The first question that the study addressed concerned the average proportion of prefabs in texts. As we can see from Table 1, which gives the proportion of slots filled with parts of prefabs in spoken and written language, the average according to our analysis is 55 percent.

<table>
<thead>
<tr>
<th>Slots</th>
<th>Filled with prefabs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>5,000</td>
</tr>
<tr>
<td>Written</td>
<td>5,246</td>
</tr>
<tr>
<td></td>
<td>10,246</td>
</tr>
</tbody>
</table>

Table 1. Proportion of prefabs in the analyzed texts

Table 1 also reveals that the density of prefabs is somewhat greater in spoken than in written language (59 versus 52 percent). This difference is not as great as expected and sometimes assumed. However, if we consider the distribution of prefab types, we find more striking differences: consider Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Lexical</th>
<th>Grammatical</th>
<th>Pragmatic</th>
<th>Reducible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>493 (38.8%)</td>
<td>261 (20.5%)</td>
<td>213 (16.7%)</td>
<td>305 (24%)</td>
</tr>
<tr>
<td>Written</td>
<td>750 (71.5%)</td>
<td>177 (16.9%)</td>
<td>25 (2.4%)</td>
<td>97 (9.2%)</td>
</tr>
</tbody>
</table>
Also, if we consider the length of prefabs, we find a difference in that prefabs are on average somewhat shorter in spoken language than in written, which we can see in Table 3.5

<table>
<thead>
<tr>
<th>Table 3. Average length of prefabs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
</tr>
<tr>
<td>Written</td>
</tr>
</tbody>
</table>

These results warrant a closer inspection of prefab types and their lengths.

4.2. Prefab types

4.2.1. Lexical prefabs
Lexical prefabs are semantic units in that they have reference and denote entities, properties, states, events, and situations of different kinds, as exemplified in Table 4.

Needless to say, notional characterizations are difficult to make precise. Here they serve the heuristic purpose of making certain that the prefab in question does represent some extralinguistic entity or phenomenon.

<table>
<thead>
<tr>
<th>Table 4. Notional categorization of lexical prefabs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties and states</td>
</tr>
<tr>
<td>out of date</td>
</tr>
<tr>
<td>be of help (to sb)</td>
</tr>
<tr>
<td>different from sth</td>
</tr>
<tr>
<td>have got sth (=possess)</td>
</tr>
<tr>
<td>be called sth</td>
</tr>
<tr>
<td>Situations and events</td>
</tr>
<tr>
<td>find one’s way (out of sth)</td>
</tr>
<tr>
<td>go out and do sth</td>
</tr>
<tr>
<td>run off</td>
</tr>
<tr>
<td>it's terrible weather</td>
</tr>
<tr>
<td>make sth sure</td>
</tr>
<tr>
<td>Places and positions</td>
</tr>
<tr>
<td>here and there</td>
</tr>
<tr>
<td>at headquarters</td>
</tr>
<tr>
<td>to the right</td>
</tr>
<tr>
<td>in industry</td>
</tr>
<tr>
<td>on paper</td>
</tr>
<tr>
<td>so far</td>
</tr>
<tr>
<td>Entities (abstract and concrete)</td>
</tr>
<tr>
<td>sketch pad</td>
</tr>
<tr>
<td>modern furniture</td>
</tr>
<tr>
<td>phone call</td>
</tr>
<tr>
<td>permanent job</td>
</tr>
<tr>
<td>subject matter</td>
</tr>
<tr>
<td>supersonic flight</td>
</tr>
<tr>
<td>intensive care</td>
</tr>
<tr>
<td>Period or point of time</td>
</tr>
<tr>
<td>at the time</td>
</tr>
<tr>
<td>by then</td>
</tr>
<tr>
<td>in the end</td>
</tr>
<tr>
<td>the eighteenth century</td>
</tr>
<tr>
<td>a long/short time</td>
</tr>
</tbody>
</table>
Indeed, trying to refer a prefab to one of the notional classes shown in Table 4, we discovered that some prefabs initially classed as lexical are difficult to place in a notional category and therefore may be on their way to acquiring functional status. Consider in reply to, of course, on these grounds, as a general thing. These were nevertheless classed as lexical. Also, included among the lexical prefabs were those with propositional content. Many of these hardly describe a state of affairs, but have a more or less clear pragmatic function: that’s true, it’s all my fault, don’t bother, is that all?, that’s a good idea, what is it all about? It is not surprising that we sometimes find that the distinction between lexical and functional items is fluid. As pointed out by construction grammarians in particular, syntactic, semantic, and often pragmatic information is connected with idiomatic phrases. Which of these features is predominant can naturally be difficult to determine. As can be seen, our policy has been to keep debatable cases in the lexical category. Admittedly a more extensive study may have justified other categories, such as routine speech acts for instance.

As the reader will already have noticed, it is possible to characterize lexical prefabs syntactically in terms of clause or phrase structure. That is to say, there are lexical noun-phrase prefabs, lexical verb-phrase prefabs, lexical prepositional prefabs, etc. Some examples are given in Table 5.

<table>
<thead>
<tr>
<th>Table 5. Syntax classification of lexical prefabs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noun phrases</strong></td>
</tr>
<tr>
<td>rule(s) of sth</td>
</tr>
<tr>
<td>sexual activity</td>
</tr>
<tr>
<td>great days of the past</td>
</tr>
<tr>
<td>maths and physics</td>
</tr>
<tr>
<td>a waste of time</td>
</tr>
<tr>
<td>the present state of our knowledge</td>
</tr>
<tr>
<td><strong>Adjective phrases</strong></td>
</tr>
<tr>
<td>ignorant of sth</td>
</tr>
<tr>
<td>able to do sth</td>
</tr>
<tr>
<td>enough of sth</td>
</tr>
<tr>
<td>all right</td>
</tr>
<tr>
<td>suitable for sth/sb</td>
</tr>
<tr>
<td><strong>Adverbial phrases</strong></td>
</tr>
<tr>
<td>once again</td>
</tr>
<tr>
<td>straight away</td>
</tr>
<tr>
<td>so far</td>
</tr>
<tr>
<td>all over the place</td>
</tr>
</tbody>
</table>
Closer inspection of the different phrases reveals typical patterns of conventionalization such as the type of prepositional link between nouns, verbs, and adjectives and their complements (*a change in sth, wait for sth, interested in sth*), prepositional verbs (*come across sth*), the type of link between verb and clausal complement (*manage to do sth, succeed in doing sth*), nominal compounds of different types (*wild animal, briefing room*), and binomials (*maths and physics*). Any teacher of English as a foreign language will recognize these patterns. But there are also preferred combinations such as *choose words* (*not pick or select*), *write a letter* (*not type*), *stick a label on sth* (*not fasten*), *switch the light on* (*not press on*), etc., etc.

The verb and the object tend to collocate more frequently than the verb and the subject.

Lexical prefabs are longer than other types of prefabs, as we can see from Table 6.

<table>
<thead>
<tr>
<th>Type</th>
<th>Average length of prefab types</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexical</td>
<td>3.03</td>
</tr>
<tr>
<td>grammatical</td>
<td>2.26</td>
</tr>
<tr>
<td>pragmatic</td>
<td>2.29</td>
</tr>
<tr>
<td>reducible</td>
<td>2.00</td>
</tr>
</tbody>
</table>

According to our mode of analysis, the length of lexical prefabs typically varies between two to five members. There are of course those that have more than five members, but they are comparatively rare.

As has already become evident, lexical prefabs are also variable as to inflection and extension and as to the type and order of members. Inflectional variability will naturally involve the choice of a determiner (*to his/her/our surprise; lay a/the table*), choice of tense (*sit down, sat down, will sit down, etc.*), aspect (*be sitting down*), and voice (*the table is laid*). The order of the members may be variable in accordance with syntactic rules: *it is going to be → is it going to be; you've got to stand up to* Hart → *Hart you've got to stand up to.*

Syntactic variability is, however, not generally predictable. For instance, one can *have a go at something, have another go at something,* but one cannot *have the go at something.* As we will see in section 5.2, restrictions of a syntactic nature upon idioms have received considerable attention from generative grammarians.

Lexical prefabs are often extendible: *due (mainly) to, lay a table (for breakfast), (close) associate of sb, (quite) all right, at (about) seven o'clock (tonight).* As has already been mentioned, if such extensions are felt to be common, they have been considered part of the prefab in question.

The most important feature of variability in a lexical prefab is the open slot. Verb-phrase prefabs will frequently have the subject slot open,
but also often the object slot: *sb look forward to sth; sb be in tune with sb/sth; sb bring sth to a conclusion; sb get in touch with sb/sth; sb miss out on sth*.

Some slots are not completely open, in which case we have what we refer to as restricted variability: *books/novels/articles etc. deal with sth; go to lectures/class/seminars/meetings etc.; have Christmas/Friday/the morning etc. off; with little/much/a lot etc. in common; to a limited/great etc. extent.*

It is interesting to note that in cases of restricted variability we find that the substitutes are normally semantically related. They are synonymous (*much, a lot*) or belong to the same lexical field (*books, articles, essays, novels, speeches, etc.*) or they are antonyms (*limited, great*). This is not an absolute rule (consider *waste of time/effort/money*), but it is certainly a clear tendency and is an indication that what we store in some cases is a meaning rather than a specific word.

We see this variability of prefabs as an outcome of two opposing forces constantly influencing language use: on the one hand there is a pull towards creating norms, facilitating communication, and at the same time there is a pull towards flexibility to allow expressibility, since expressibility increases with nonpredictability.

### 4.2.2. Grammatical prefabs

Grammatical prefabs are intralinguistic text-forming items rather than units with extralinguistic reference. In our classification there are nine types of function:

1. determiners: *(the) sort of, of some kind, (the) kind of, the same, the next, the rest of, my own, the right;*
2. quantifiers: *a bit (of), one of, a little (of), a little bit (of), most of, more of, a lot (of), an amount of, part of, all parts of;*
3. proforms: *each other, anyone else, no one, things like that, the kind of things, that which, or something, all this, one of them, most of them, a few more, some people, down here, up here;*
4. introductors (existentials and identifiers): *there is ..., this is ..., that is ... ;*
5. tense-forming: *be going to, have been;*
6. aspect-forming: *begin to, be about to, tend to;*
7. mood-forming: *may be, might be, can be, could be (possibility); have to, have got to (necessity); seem to, must have been (epistemological); would have been (hypothetical); ought to (deontic); let us (hortative); had/would rather (optative); ?if necessary (conditional);*
8. links: as well (as), and so, as yet, as if, on the other hand, now that, either ... or, so ... that, as regards, apart from, out of, in front of, in spite of, among other things, more than, as much as;

9. intensifiers: so very, very much, much less, more or less, even more, not at all, not in the slightest, not even, at least, if anything.

This classification needs some explanation:

i. We have argued that the same, the next, my own, etc. form determin- ing prefabs since the choice of one of these restricts the choice of the other (*a same, *a next, *the own are not possible).

ii. There is some overlap in the classification in that many of the quantifiers are types of determiners, but since not all of them are, they have been allowed to form a separate group.

iii. Links include conjunctions and adverbs as well as prepositions and also comparatives such as more than and as much as.

iv. Among the intensifiers we find intensified negations (not at all) and one downtoner (not very). There are some candidates for quantifier status which we have left among lexical prefabs, viz, a drink of, a cup of, a fistful of, a spoonful of.

The list of grammatical functions reveals our justification for considering grammatical prefabs as text-forming rather than referring units. That is to say, a grammatical prefab will quantify, specify, modify the reference or meaning of nouns, verbs, adjectives, or adverbs, in a general manner, or they will serve as their substitutes or as links between propositional or referring items.

Structurally, grammatical prefabs differ from lexical prefabs in that they are normally shorter (typically varying between two to four members) and not as variable as lexical prefabs. It is true that some of them are inflectionally variable and that some may be extended (a [great] amount of, [not nearly] as much as) and that some allow a restricted choice of members (a great/vast/awful amount of), but this variability is much more limited than in the case of lexical prefabs and many grammatical prefabs are quite frozen.

Arguably some grammatical prefabs involve open slots: links must link something, determiners must have heads, introductors introduce some- thing, etc. We decided against marking open slots of this kind, i.e., open slots which by definition are connected with certain grammatical prefabs. They are different from the open slots in lexical prefabs. Also, marking them would force us to italicize great amounts of text obscuring the notation. This policy does not affect the statistics, since in calculating number of slots filled with parts of prefabs, words in open slots are ignored anyway.
4.2.3. **Pragmatic prefabs**

Pragmatic prefabs are functional in that they do not directly partake in the propositional content of the utterance in question. They differ from grammatical prefabs in that they may occur outside the syntactic structure. Most of them are restricted to spoken language and some have functions which could be indicated by punctuation, paragraphing, or in other graphic ways in written texts.

Brinton (1996: 38) suggests that there are two main categories of pragmatic markers; viz. textual and interpersonal. We have found it difficult to do without a third, i.e., those with metalinguistic functions. Our classification is therefore along the following lines:

1. **Text monitors:**
   - discourse markers (*and then*)
   - turn regulators (*well you know*)
   - repair markers (*I mean*)

2. **Social monitors:**
   - interactives (*wouldn’t it*)
   - feedback signals (*I see*)
   - hesitations (*what’s the word*)
   - responses (*yes I think so*)
   - performatives (*do sit down*)

3. **Metalinguistic monitors:**
   - approximators (*and everything*)
   - hedges (*sort of*)
   - epistemological signals (*I should think*)
   - attitudinal markers (*I must say; my dear*)

Discourse markers are text-oriented markers which are used as indicators of various kinds of transitions in discourse, for instance marking boundaries between topics, between modes of speech (direct and indirect speech), between foregrounded and backgrounded information in the thematic structure, and as cohesive devices between sets of propositions at a textual level, comparable in function to conjunctions at the sentence level (cf. Schiffrin 1987; Fraser 1990; Redeker 1990). Typical examples include *and then, and finally, and of course, but anyway, the thing is that ..., you know, I mean, and so on, well I thought, she said, as I said*.

Turn regulators are text monitors marking transitions between contributions made by different speakers. They also serve to regulate who speaks and who listens, and above all, to keep the channel open between speakers and listeners (Sacks et al. 1974; Pomerantz and Fehr 1997). Examples in this group are *you see, you know, well I think, well you know*.
Repair markers are text monitors functioning in the editing of the discourse. The speaker may wish to change the wording, repair a mistake, or give more adequate information related to the previous discourse; whatever the reason, this usually brings about a break in the encoding of the message, or a deviation from the ongoing line of argumentation. Repair markers include *I mean, you know, well you know.*

Interactives are classified as social monitors in that they elicit audience involvement by calling for action on the part of the addressee, e.g., confirmation of a previous claim. The most obvious instances include tags (*wouldn't it?, has he?*) and equivalent elicitors, e.g., *you see what I mean, you know, you see,* the latter two usually with a rising tone.

Feedback signals serve to confirm audience involvement by ensuring that the channel is open between speaker and addressee(s). They are often referred to as back-channel signals. Examples include *well no, I see, suppose so, has he,* usually with a falling tone.

Hesitation markers have a social function because they are signals to the addressee that the speaker wishes to continue in spite of some encoding difficulties, either in retrieving a particular word or phrase from memory or in overall planning of the continuation of the utterance. A typical prefab with this function is *what's the word.* But *you know or I mean* are also frequently used to stall for time.

Responses are like feedback signals in that they confirm audience participation, but unlike feedback in that they constitute a turn and are answers to a question. Examples are *oh no, well yes, yes I see.*

In all these cases we can see that pragmatic prefabs are natural developments given the nature of conversation. In conversation we have to be open about our errors; turn-switching is an art that may need prompters; there are no punctuation marks or paragraphing to indicate topic shifts. We expect constant confirmation that our message is being processed.

Performative routines (*thank you, good luck, why don't you ..., good evening,* etc.) have a very different function from the pragmatic prefabs just mentioned in that they elicit or constitute actual performances such as thanking, making offers, or leave-taking.

The metalinguistic comments, in particular the hedges and the epistemological prefabs, also form a special category. They serve to relieve the speaker from being completely committed to the truth value of the proposition in question. These prefabs have counterparts in written texts, i.e., frequently modal auxiliaries and adverbs (*perhaps, probably, in all likelihood*). Hedges and epistemological prefabs are close. Examples of hedges include *or something, or somewhere, and all this,* and of epistemological prefabs *I should think, I dare say, I must say.* Some
attitudinal markers are responses rather than metalinguistic in nature, notably my dear, by golly.

Discourse analysts working with pragmatic markers often point out that these tend to be multifunctional, i.e., one and the same marker can have more than one function, not only in different contexts but also in one and the same context. Not surprisingly, we have had the same experience. The markers you know, you see, and I mean, for example, may serve as social monitors and text monitors simultaneously. However, as all discourse analysts are well aware, one function is usually predominant over the others, as made evident through particular contextual features.

The tendency to be multifunctional is a feature that pragmatic prefabs do not share with grammatical prefabs, at least not to the same extent. It would be difficult to find proforms which could also function as modal auxiliaries, or links which could also function as determiners. They also differ from grammatical prefabs in that they occur almost exclusively in speech (or reported speech). They are like grammatical prefabs in that they tend to be short and relatively invariable and appear to be restricted in number. This is reflected in token–type ratios. In the case of lexical prefabs the token–type ratio is 1000–980; in the case of grammatical and pragmatic prefabs 1000–650 and 100–660, respectively.

4.2.4 Reducibles

Reducibles in our texts fall into the following four main groups:

i. Pronouns + am/is/are/have/shall/will, e.g., I'm, it's, they've, you've
   he'd, I'll

ii. Auxiliaries + not, e.g., don't, isn't, hasn't, can't, shouldn't, wouldn't, needn't

iii. Auxiliary + auxiliary, e.g., would've, should've

iv. let + us: let's.

Although reducibles consist of grammatical words, they have no common grammatical function. They are formed because, given the context, the reduced element is predictable and reduction is phonotactically possible. The general trend of short and unobtrusive function words allowing content words to have prominence is possibly an important impetus for their formation.

Occasionally we find that reducibles override inflectional and syntactic rules: how's things, I'm English aren't I, let's us go there, let's you come here (American English). We take this as evidence that reducibles are stored as ready-made items.

Returning to Table 2, we see that reducibles are significantly more common in spoken than in written language. It is possible that we may
have overlooked some reducibles in our analysis of written texts (where they are written out in full), but such occasional omissions cannot account for this great difference. A more plausible explanation is that pronouns and tag questions involving *not*-reductions are more common in conversation than in written texts.

4.3. *Combinability*

Prefabs may conjoin with each other or some other word or words to form a single *composite prefab*: *a little bit* + *more than*; *that one* + *over there*; *due to* + *the fact that*; *wonder* + *what to do*; *switch on* + *the light*; *books* + *deal with* + *some topic*.

As has already been exemplified, a prefab may be embedded in another prefab. (We remind the reader that italics indicate words in open slots and that embedded prefabs are within square brackets.) It is in fact fairly common for a prefab to fill the open slot of another prefab:

(8) /[[the Prime Minister] and Mr Lloyd had a (whispered) conversation/ /they nodded at [each other]]/ [LOB:A06 005–152]
(9) /the preparations for [launching their rockets]/ [LOB:A06 005–152]
(10) /the average of [forty miles [an hour]]/ [LOB:N01 002–103]

A prefab filling an open slot may in turn have an open slot. This we refer to as *double open*:

(11) /[[the first] hint of [the Chancellor bowing to [public opinion]]]/ double open [LOB:A06 005–152]

(11) is an example of the textual complexity which can arise in combining prefabs. Here we have a prefab containing two separately embedded prefabs, one of which contains yet another prefab.

There are a fair number of apparent *overlaps*, i.e., the end of one prefab is simultaneously the beginning of another.⁷

(12) /Hart [you['VE] got to] stand up to /haven’t you/ [LLC:1.5.117]

'VE is the end member of a reducible and forms at the same time the beginning of the grammatical prefab *have got to*. All our examples of overlaps involve reducibles. They occur when a prefab begins with a reducible item (e.g., *have in have got to*) and when this prefab is preceded
by a word which makes reduction possible (e.g., you). Since reducibles do
not form semantic or grammatical units, we suggest that these overlaps are
apparent. It is as if the first member of a reducible is pulled into a prefab
and made part of this by its partner (which is the first member of the
attracting prefab).

Prefabs may of course occur successively:

(13) A /I gather/ /[you’ve] been at it/ /for nine years/
a /By golly/ /[that’s] true/
[LLC.2.2a]

(14) /Just as/ /she was wondering what to do (next)/ /she came upon the
Three Bears’ little house./

But often prefab and non-prefabricated strings are interspersed:

(15) /Just as/ /she was wondering what to do (next)/ /she came upon the
Three Bears’ little house./— — — — — — — /how nice it would be to
have someone to say, /‘Come in, /my dear. /Sit down—/ have some
breakfast.’—/no one/— — — — /in answer to her gentle knock/—/she
ran round/—/the window peeped through/— — — — —
[From The Three Bears, Jane Carrath, London: Cathay Books, 1982.]

(16) A /I gather/ /[you’ve] been at it//for nine years/
a /By golly/ /[that’s] true— — /it’s— /a long time//of course/
(in the) in this ([sort of]) work/ /you know/
A /well no— /it’s— /a long time/ <syll> > /by any standards/
a — /suppose so/ <1 syll>
[LLC.2.2a]

In our material the length of non-prefabricated strings varied from
one to eighteen members, if we consider as non-prefabs only those we
have indicated by dashes. Not surprisingly, the non-prefab strings tended
to be longer in written than in spoken language. In spoken language the
longest non-prefabricated string consisted of 11 slots. An example of
a long non-prefab string (14 slots) is the following:

(17) — — — — — — — — — — — — — /over to the table/—/sat
down/ /in front of [the biggest bowl of porridge]./

Full text:
The lovely creamy porridge made Goldilocks remember that she
was very hungry. She skipped over to the table and sat down in
front of the biggest bowl of porridge.
[From The Three Bears, Jane Carrath, London: Cathay Books, 1982.]
The great majority of non-prefab strings consist of one or two members. (This trend is particularly strong in spoken language.) Non-prefab strings of more than ten slots are rare. Generally it is true to say that the longer the non-prefab string, the rarer it is. Table 7 gives the average lengths of non-prefab strings.

Table 7. *Average length of non-prefab and prefab strings*

<table>
<thead>
<tr>
<th></th>
<th>Non-prefab</th>
<th>Prefab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>2.47/2.16</td>
<td>2.61/4.01</td>
</tr>
<tr>
<td>Written</td>
<td>3.84/3.02</td>
<td>2.80/4.08</td>
</tr>
</tbody>
</table>

(The first figure under *non-prefab* refers to non-prefab strings indicated by dashes; the second to non-prefab strings including also adjacent open slots and open slots occurring within prefabs, which tend to be short, thus explaining the lower average figure. The first figure under *prefab* refers to single prefabs; the second to strings of successive prefabs.)

The figures in Table 7 show that both prefab and non-prefab strings are longer in written than in spoken language, the effect being particularly significant in non-prefabricated strings. In fact, of the differences between spoken and written language that our study revealed, the most important is that in written language we have longer non-prefab strings.

### 4.4. Number of choices

Since practically all texts contain prefabs and since prefabs can be assumed to constitute single multi-word retrievals from our mental store of words, the number of retrievals must be fewer than the number of words in a text of some size. However, as we have seen, prefabs are variable to some extent. Therefore determining the number of prefabs in a text does not necessarily reveal the number of retrievals. Consequently we had to make a separate analysis of the number of choices in a text.

However, what represents a choice is far from self-evident. Consider the following five-word sentence:

(18) She stands up to him.

At first sight, it seems reasonable to postulate that it represents three choices: *she* + *stands up to* + *him*. But, arguably, it also involves choosing the nominative form for the subject, the accusative form for the object and the present tense for the verb. We have, however, decided to consider only choices of slot-fillers (i.e., words). Expressed differently, in our analysis one slot can never represent more than one choice. Since choice of tense sometimes involves two slots in English we decided that, for the
sake of consistency, tense-forming auxiliaries do not represent separate choices. (This has already been pointed out earlier in the article.)

Let us consider another problem, exemplified by the following sentence:

(19) She will stand up to him for a short time.

Since there is evidence that for a short time is a prefab, we could posit that this ten-word sentence involves four choices: She + will stand up to + him + for a short time. However, the prefab for a short time involves restricted variability: for some / a short / a long time. Should we therefore posit that this prefab involves yet another choice (for a + short + time), making the total number of choices five? We decided to do so. This implies the belief that the language user has chosen a particular prefab construction and, of the possible members, the one that matches the intended message. We admit, however, that it could be argued that for short time constitutes a single choice, representing a single notion corresponding to 'briefly'.

In view of problems of this kind, we cannot claim that our estimations of choices involved in the texts we analyzed are indisputable. We do think, however, that they give some idea of processing effort involved in composing a text. In any case they allow comparison between text types in this respect. Table 8 compares number of choices in spoken and written language.

Table 8. Number of choices in spoken and written language

<table>
<thead>
<tr>
<th>Spoken</th>
<th>Written</th>
</tr>
</thead>
<tbody>
<tr>
<td>68%</td>
<td>75%</td>
</tr>
</tbody>
</table>

As expected, there are more choices made in composing a written than a spoken text. This difference is statistically significant, $p < 0.001$ as evidenced through both 'goodness-of-fit' and 'contingency table' testing. Moreover there is a lot more repetition in spoken texts, which these figures do not show.

Presumably, choosing an item from a restricted set is less demanding than selecting one from an indefinitely large set. Therefore we also estimated how many of the choices were lexical in nature. Table 9 displays the results of this analysis of our spoken and written samples.

Table 9. Number of lexical choices in spoken and written language

<table>
<thead>
<tr>
<th>Spoken</th>
<th>Written</th>
</tr>
</thead>
<tbody>
<tr>
<td>27% of all slots</td>
<td>41% of all slots</td>
</tr>
<tr>
<td>39% of all choices$^a$</td>
<td>55% of all choices</td>
</tr>
</tbody>
</table>

$^a$For clarification of this distinction, see the sample text in the Appendix.
Again, not unexpectedly, there are more lexical choices in written than in spoken language. Needless to say, this difference is also statistically significant, $p < 0.001$ as evidenced through both 'goodness-of-fit' and 'contingency table' testing.

As an illustration of our analysis of choices, an extract from one of our written text samples is given below. Each postulated choice is indicated by an unbroken line.

(20) Lois's arrival seemed to warm and enliven the atmosphere, and, as Joan had predicted, there were no gaps in the conversation—She was not in the least self-conscious and so obviously bubbling over with youthful high spirits that the two men could be almost seen to thaw in her presence. Before she came there had been a slight stiffness, due chiefly to the fact that they were practically strangers with little in common. 50 choices, 73 slots.

4.5. Summing up the results

The first question the present study set out to answer concerned the average proportion of prefabs and average number of choices in a text. We have found that on average somewhat more than half (around 55 percent) of a text will consist of prefabricated language. This suggests that in a text of 100 words on average only 45 single-word choices would be made. This figure does not tell us how many choices there are in total in a text. A separate analysis of choices gave us the average figure of 71 choices in 100 words. If we compare the figure for single-word choices (45) with the figure for the total number of choices (71), we realize that there are only 26 choices instead of the 55 that there would have been had there been no prefabs. Of these 26, quite a number constitute restricted choices within prefabs requiring comparatively little mental effort. Even though these figures are approximate (in view of the fact that it is far from indisputable what constitutes a choice or a prefab), they nevertheless are a measure of the high degree to which preconstructed language facilitates the production and presumably the interpretation of utterances. They give strong support to the idiom principle as formulated by Sinclair (see the Introduction) and reveal that the proportion of prefabrication in language has generally been much underestimated.

As far as the second question raised in section 1.2 is concerned, i.e., whether the frequency of prefabs depends upon whether the text is
spoken or written, we found that conversation exhibits a greater proportion of prefabricated language than does written language, but that the difference is not as great as expected. The striking difference between spoken and written language concerned the distribution of prefab types and the length of non-prefabricated strings.

As far as combinability is concerned, we found that prefabs can be conjoined with each other and with other words to form larger prefabs, that they can be embedded within one another, and that there are successive sequences of prefabs. Above all, we found no example of a text that consisted exclusively of prefabs or which contained no prefabs. This supports our basic hypothesis, namely that the production of a text involves frequent alternation between prefab and non-prefabricated strings.

As to the length of prefabs, we found that the average length of single prefabs (as opposed to prefab strings) is—perhaps surprisingly—modest (2.7 words). It is important to keep in mind, however, that their length depends on type of prefab, in particular whether they are lexical or not. With regard to the length of non-prefabricated strings, we found that they tended to be somewhat longer on average than prefabs. Not surprisingly, variation in length is greater in non-prefab strings than in prefabs. Further, in written language their length is on average greater than in spoken language. The reader is reminded, however, that the most frequent length for non-prefabricated strings in the written medium as well as in speech is one to two slots.

As could be expected, there are different types of prefabs. We found the following classification feasible: lexical prefabs, which are semantic units; grammatical prefabs, which are text-forming units; pragmatic prefabs, many of which are text-monitoring but which do not replace but rather complement grammatical items and which tend to be restricted to the spoken medium; and finally reducibles, which are types of abbreviations. This classification is not without problems, in particular we have occasionally found it difficult to determine whether a prefab is grammatical or lexical and sometimes whether it is grammatical or pragmatic. Moreover it is possible that a greater number of prefabs would cause the formation of a class of conventionalized speech act prefabs (so called routines) which would be clausal in structure and pragmatic in function, but not necessarily restricted to the spoken medium.

Finally, we wanted to establish the variability of prefabs. Unlike words, prefabs have syntactic structures. Therefore it is not surprising that many prefabs allow inflectional variability and allow the order of the members to be changed according to syntactic rules. They are unlike non-prefabricated syntactic structures in that their syntactic flexibility is
restricted in what seems to be a non-generalizable manner. Prefabs may also be semantically flexible. They may allow insertions modifying certain members ([about] an hour and a half, a [big] spoonful of sth, the goings on [of the time]). Above all, members may be substitutable, either predictably (restricted variability) or without restriction (open slots). Generally, restricted variability occurs when a particular slot has to be filled not with a particular word but with any word representing a particular kind of entity or phenomenon. Consider this example: books, novels, articles, speeches, talks, etc. deal with some topic. Open slots are common with verb-phrase prefabs, which follows from the fact that they have argument structures just like non-prefabricated verb phrases. However, open slots can also be modifiers: for some purposes, at a sth time. Variability is more common with lexical than with functional prefabs. This is natural since variability is more likely to enhance the usefulness of lexical prefabs than that of grammatical prefabs. Finally, it should be pointed out that syntactic as well as semantic flexibility depends on transparency. That is to say, if a multi-word combination is opaque so that we cannot see how the different members contribute to the meaning of the whole, we tend to be at a loss as to how to manipulate them. This is in essence the point made by Nunberg, Sag, and Wasow (1994) and further elaborated in Hudson (1998).

The foregoing represents our account of the present study. We will conclude by considering first (and very briefly) some practical and then some theoretical implications of our findings.

5. Implications of the study

5.1. Practical implications

The recognition of prefabs and their importance in communication together with accurate descriptions of their meanings and/or functions, their syntactic and lexical flexibility and their style level would have far-reaching practical implications for a number of areas, notably:

i. Foreign language learning and teaching. Raised awareness of the abundance and importance of prefabs should improve the student’s learning strategies and command of a foreign language. Teaching material could probably be adapted to better represent and highlight nativelike use of language.

ii. Machine translation. If it were possible to build up databases of prefabs contrastively, machine assisted translation would most probably be considerably more efficient than it is today.

iii. Genre studies. These would benefit significantly if we were able to specify common core and genre-related prefabs.
5.2. Theoretical implications

In the last three decades or so idioms have attracted the attention of linguists for different reasons. Generative grammarians have worried about idioms because they do not conform to the assumption that lexical insertion rules introduce single words—i.e., items with no syntactic structure—into phrase structures. Moreover, idioms have non-compositional meanings and do not allow syntactic manipulations in a predictable fashion, which make them difficult to account for by means of rules. A number of solutions have been suggested. Weinreich (1969), Newmeyer (1974), Fraser (1970), and Chomsky (1980) represent the best known early attempts. Nunberg, Sag, and Wasow (1994) and Jackendoff (1997), and above all Construction Grammar (Fillmore et al. 1988; Kay 1997), represent some well-known contemporary attempts. Fillmore, Kay, and O'Connor (1988) and Jackendoff (1997) both point out that idiomaticity is not a marginal phenomenon. The former (1988: 504) suggest that the repository of what is idiomatic in language is very large and Jackendoff (1997: 156) estimates that the number of memorized fixed expressions is of about the same magnitude as that of the single words of the vocabulary. The fact that between 40 and 60 percent of the 19 text excerpts analyzed in this study consisted of ready-made and to a large extent idiosyncratic combinations of words indicate that these surmises are correct. We therefore join the growing number of linguists who abandon the traditional view that in encoding and decoding we draw upon two distinct types of linguistic knowledge: the grammar, supplying us with the knowledge of how to inflect and combine words, and the lexicon, supplying us with knowledge of the form and meaning of words, and which together make it possible to produce and work out the meanings of phrases and sentences—a model outlined in Figure 1.

![Figure 1. Traditional model of compositionality](image)

Instead, we envisage a model which makes allowances for the observation made by so many contemporary semanticists, viz. the meanings of words are negotiable, and context and shared background knowledge
Britt Erman and Beatrice Warren

play an important part in interpretation. As an example of this, let us consider the different interpretations of write in some contexts:

(21) a. write a greeting on a Christmas card
    b. write a check
    c. write a letter
    d. write slogans on walls
    e. write some words in the sand

In (a) and (b) we imagine a pen or pencil as the instrument in (c) we envisage no particular tool, it could be a typewriter, a word processor as well as a pen or pencil, in (d) the instrument could be a paintbrush; in (e) it is likely to be a stick or a finger. The different interpretations are produced in matching the phrases with plausible referents. We assume that there can be no interpretation without such a match and that matching necessarily involves consulting our knowledge of the world and the context at hand, if there is one. If there is not, we will create a kind of default context. We see the contribution of context and knowledge of the world to the interpretation of linguistic units as the source of non-compositionality. Sometimes the modification of conventional meanings may be so drastic that linguists will think of them as figures of speech, but note that this is far from always the case, as the examples in (21) demonstrate. Reading literature on phraseology one gets the impression that for anything to be non-compositional, it should be opaque or syntactically irregular or involve words used figuratively. In our view syntactic or semantic irregularity makes the non-compositional character of a phrase evident, but it is a mistake to believe that whatever is transparent is also compositional. It is probably because of this assumed connection between transparency and compositionality that idiomaticity has been looked upon as a marginal phenomenon. Indeed, in the written sample text in the appendix, most linguists would probably find only three idioms, i.e., the ones underlined in the full written version of text G77 in Appendix 2.

As exemplified in (21), phrases and their interpretation may stick and become conventionalized, in particular if they denote some repeated situation. Recall Pawley and Syder’s (1983) point that idiomaticity arises because standard situations naturally tend to be expressed in standard ways. Of the combinations in (21a–e), write a letter and write a check represent common and fairly important activities, which may explain why they meet our criteria for prefab status. (Write a check is clearly not compositional—at least not in our understanding of the term. When one writes a check, one does not write the word check, one writes something rather specific on the check.) The contribution of context and shared
knowledge of the world to the interpretation and eventually to the meaning of words can also explain the origin of opaqueness in idioms. Originally idioms will naturally have been transparent given shared background and contextual knowledge. Opaqueess arises when contextual and background knowledge are no longer available.

In this connection we may remind the reader that we had observed that it was more common for the grammatical object to be part of a verb-phrase prefab than for the subject and the verb phrase to form a prefab. Similar observations have been made by others and are discussed in Nunberg, Sag, and Wasow (1994: 525–531) in the case of idioms. These authors suggest that the fact that subjects are relatively rarely parts of idioms is due to the fact that AGENTS are characteristically animates, and animates are relatively rarely affected by metaphorization. An alternative explanation is that a process affecting or effecting some entity is more likely to influence the character of this entity than the perpetrator.

The model we suggest should replace the traditional model is outlined in Figure 2. In Figure 3, we then complete the revision of the model of knowledge of language. That is to say, we postulate that memorized, lexically specified knowledge consists of single words (the lexicon) and
the phrasicon (prefabs).\textsuperscript{10} In producing and interpreting utterances (combinations of words), we normally draw upon the lexicon, the phrasicon, grammatical knowledge, knowledge of the world, and the context at hand. Most of the utterances we decipher and form will be discarded,\textsuperscript{11} but occasionally bits of varying size will be retained. Some of these will be additions to the lexicon, others will be additions to the phrasicon. Additions to the lexicon will be new senses of established single words or new derivatives. Additions to the phrasicon will be prefabs, i.e., combinations of words with (at least initially)\textsuperscript{12} transparent syntactic-semantic structures, and therefore keep their links to the grammar and the lexicon. (This is indicated by broken lines.) These links make prefabs syntactically and sometimes semantically flexible in ways which are not possible in the case of items in the lexicon.

The reader is warned that the model in Figure 3 is a rough sketch rather than a finished design. The distinction between functional and lexical words and phrases is ignored. Nor are the grammaticalization process or word-formation rules accounted for. The main purposes of the figure are to highlight (a) that the source of polysemy and non-compositionality is the constant matching and assimilation of linguistic with nonlinguistic knowledge involved in interpretation, and (b) that stored lexical knowledge is of two distinct types. We can no longer look upon the mental lexicon as a store of single words with the odd idiom thrown in. However, the boxes in the figure should not be taken to imply that grammar, the lexicon and phrasicon are rigidly delimited components of linguistic knowledge. Instead we envisage a cline going from what is fixed and unanalyzable to patterns allowing a great deal of variation, i.e., from single items through fixed phrases, variable phrases, formal idioms to basic sentence patterns such as intransitive, transitive, and ditransitive structures, etc.\textsuperscript{13}

The addition of a phrasicon to the lexicon would bring about something of a revolution in semantics in that it should considerably influence our understanding of concept formation. The flexibility of concepts will probably no longer be seen as restricted to grammaticalized phenomena such as tense and number and mood, but as possibly involving an indefinite number of participating entities reflected in the open slots of prefabs, or being restricted to some types as reflected in the restricted variability of some prefabs. Such a view would support the position of current cognitive theories, i.e., that categories are non-static and meanings have an experiential basis.

Finally, it is obvious that prefabricated language is relevant not only to theories of syntax and semantics but also to psycholinguistics. Psycholinguistic research into the processing of idioms has shown that
comprehension of idioms is faster than comprehension of phrases composed according to the open-choice strategy (for an overview of this, see Peterson and Burgess 1993). Presumably this is true of prefabs too and applies not only to decoding but also to encoding effort. The attempt of the present study to establish lengths of prefab and non-prefab strings and the number of choices made in composing a text—whether written or spoken—reflects our view that estimating encoding and decoding effort for texts should involve all types of prefabricated language, not only bona fide idioms and that a method might be worked out that measures processing effort fairly accurately.

Appendix 1: Terminology

Composite prefab. A composite prefab will consist of at least one prefab conjoined with another prefab or word. For examples, see p. 104.

Dictionary. The list of prefabs found in a text excerpt that has been analyzed. See Appendix 2 for an illustration.

Discontinuous prefab. A prefab the members of which are not adjacent. In the prefab analysis an asterisk will follow the first and precede the second member: so* *that.

Double open. A prefab containing an open slot filled with a prefab which in turn contains an open slot. Examples are given on p. 104.

Extension. Non-obligatory parts of prefabs. These are enclosed within parentheses.

Non-prefab string. Combinations of words in a text which have been put together according to the open choice principle. In the prefab analysis words combined like this are replaced by dashes.

Open slot. A part of a prefab which can be filled in a practically unlimited way by lexical material, but which must be filled for the prefab to be complete. In the dictionary such members are indicated by some, somebody, something. In the prefab analysis they are italicized and reduced in size to indicate that they are not counted as representing preconstructed language.

Overlap. We have an overlap, if the last member of a prefab is simultaneously the first member of a following prefab. We have found that overlaps invariably involve reducibles. That is to say, the overlap is always the last member of a reducible. Therefore we suggested that there are no true overlaps. See p. 104. The overlap is marked with capitals in the prefab analysis.

Prefab within prefab. If an open slot is filled with a prefab, we have a prefab within a prefab. Such prefabs are italicized and enclosed within square brackets. We also have embedding if an extension of a prefab
Prefab string. A prefab string can be a single prefab or two or more successive prefabs uninterrupted by non-prefabs. Restricted exchangeability. The criterion for prefab status made use of in the study. If normal syntactic or semantic variability is not possible without a change of meaning or function and/or idiomaticity, we have restricted exchangeability. See pp. 90–91. Restricted variability. If a member of a prefab can be replaced in a limited manner, we refer to that as restricted variability. Variable members of this kind are not marked in the prefab analysis, but in the dictionary they are indicated as follows: to some/a large/no etc. extent. Slot. Each orthographic word in a text is considered to fill a slot.

Appendix 2: Sample texts

G 77 001–004
To the best of my knowledge, there is no record of a society which has used literacy for the profane and imaginative purposes and which has not produced books dealing with sexual topics; (33 slots)

/To the best of my knowledge/ /there is [no record of a society]!/— — — — /for the profane and imaginative purposes/ /has not /— /books dealing with sexual topics;/ (20)

To the best of + my + knowledge, there is no record of a society which has used literacy for + the profane and imaginative + purposes and which has + not + produced books dealing with sexual topics; (22/13)

33 slots
20 filled with portions of prefabs
23 choices (69.7% of the slots)
13 lexical choices (39.4% of the slots; 56.5% of the choices)
4 + 1 + 1 = 6 prefabs

Lexical
to the best of my/our etc. knowledge
a record of sth
for the profane etc. purposes
books /articles etc. deal with sexual topics/matters etc.

Grammatical
there is (exist) NB reducible
Reducible

To the best of my knowledge, there is no record of a society which has used literacy for the profane and imaginative purposes and which has not produced books dealing with sexual topics; of these books some have been considered unsuitable for general reading, their circulation has been more or less clandestine, and where laws have been concerned with private morals have been interdicted by the law. As far as I know, there is no surviving pornography from Mesopotamia, Pharaonic Egypt or Crete; but there is so little written matter surviving from these civilizations that no argument can be based on these omissions.

LLC 1.05

A isn't this going to be a strange and impossible task for me picking up linguistics and I'm entirely at least almost entirely ignorant of it {at present}
B < 2 sylls> #
A < approx 4 sylls> #
B go to find out the right seminars {to go to} that's what I did when I first came
A the right seminars yes#
B < you know > I went to some [s]
A cos *lectures are rather a waste of* time < are they >

A /[is*n't] this *going to be — — — — — — — — — — — — — /picking
up linguistics/ — /I'm (entirely) ([at
least]) (almost entirely) ignorant of it
/at present/
B < 2 sylls > #
A < approx 4 sylls > #
B — — /find out [[the right]
seminars to go to]/ /that's/ — — — — — — — — — — — — —
A /[the right] seminars/ —
B /you know/ /I went to some [s]
A — — — — /a waste of time/
/are they/
Britt Erman and Beatrice Warren

B *what did you read English* or not#
A yes I read English but only from Kyd onwards so that you see didn’t even do any {Old English or any Anglo-Saxon at all}

61%

Notes

1. This lecture was later published (Bolinger 1976).
2. In this connection, it should be pointed out that we have not considered proper names as prefabs although these are often memorized combinations of words. Our reason is that names are not dictionary items and arguably not part of langue.
3. Terms used in the study are italicized when first introduced. We supply a complete list of terms in Appendix 1.
4. Either ... or is an example of a discontinuous prefab. In the prefab analysis these are marked in the following way: either* ... *or.
5. These averages are somewhat uncertain since they have been calculated on the basis of prefabs from text extracts of 1000 words for each category. We could not estimate average lengths by dividing the number of slots filled with prefabs by the number of prefabs, since in the dictionary we listed all prefabs even those which were part of other prefabs. For instance, run in on engine power (about aeroplanes) was listed under run in on engine power, run in, on engine power, engine power.
6. In accordance with lexicographic tradition, if the subject slot of a verb-phrase prefab is open, this has not been indicated in the dictionary—a practice we also adhere to elsewhere in the article.
7. Overlaps are marked by capitals.
9. Originally it would have been a pen, but since what is perceived to be important about this activity is the conveyance of a message on paper, a process in which meaning adapts itself to salient aspects of reference has effected the change in question. This type of process was formerly termed adequation by early semanticists. See Stöcklein (1898) and Stern (1965).
10. The term phrasicon is adopted from Fillmore, Kay, and O’Connor (1988).
11. This is indicated by the dotted line at the bottom of Figure 3.
12. In the case of learners of a language, the situation may be the reverse: links may not exist at first but become established as the lexicogrammatical structure is worked out. (See Wong-Fillmore 1976, and Peters 1983).
13. By formal idioms we mean the types of idiosyncratic syntactic structures identified by Fillmore, Kay, and O’Connor (1988: 510–511) and exemplified by Him be a doctor!

References

Britt Erman and Beatrice Warren


Britt Erman is lecturer at the Department of English, Stockholm University, Sweden. Her research interests are pragmatics and discourse analysis. She is the author of Pragmatic Expressions in English (1987).

Beatrice Warren is professor of the English language at Lund University, Sweden. Her main research interests include word formation and lexical semantics, which is reflected in the three monographs she has written, Semantic Patterns of Noun–Noun Compounds (1978), Classifying Adjectives (1984), and Sense Developments (1992).