Register Differences between Prefabs in Native and EFL English

MARIA WIKTORSSON

1 Introduction

In the later stages of EFL (English as a Foreign Language) learning, and foreign language learning in general, emphasis is put on approaching a native-like command of the target language. Native-like command of a language entails, besides using correct grammar and vocabulary, being able to use the language idiomatically. However, idiomaticity is not an easy concept to define, or as Yorio (1989) puts it:

> Idiomaticity is a non-phonological "accent", not always attributable to surface language errors, but to a certain undefined quality which many frustrated L2 composition teachers define as "I don't know what's wrong with this, but we just don't say that in English". (Yorio 1989:64)

If we, as researchers of language learning and teachers of foreign languages, do not want to content ourselves with the "I don't know what's wrong with this"-approach we need to find ways of investigating our students' target language idiomaticity.

In this study it is proposed that one way of investigating the idiomatic nature of language is by means of prefab-identification and analysis. Prefabs can be defined as follows:

> a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray & Perkins 2000:1)

Previous research on prefabs in essays written by advanced Swedish EFL learners and native speakers (Wiktorsson 2000), showed quantitative differences between the types of prefabs used. It was hypothesised that these differences were the result of a lacking register awareness on the part of the learners.

The present study will further investigate that hypothesis by selecting a subset of the prefabs found by Wiktorsson (2000). These will be checked in two corpora, one of speech and one of writing, in order to find out if the learners' prefabs are more frequent in speech.

1.1 Theoretical background

Several researchers over the past couple of decades have begun to question the traditional approach to language production, i.e. that a language user creatively produces language by combining single lexical items from the mental lexicon according to certain syntactic rules. For instance, Pawley & Syder (1983) claim that this approach cannot account for native-like selection and fluency since:

> only a small proportion of the total set of grammatical sentences are native-like in form – in the sense of being readily acceptable to native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', 'odd' or 'foreignisms'. (Pawley & Syder 1983:193)

As an illustration of the above we can use sentences 1 and 2:

(1) My name is Maria.

(2) I have the name Maria.

Sentence (1) is the natural native English way of introducing oneself, but (2), which expresses the same thing and is equally grammatically correct, sounds odd.

The above difference between only grammatically correct and grammatically correct as well as preferred ways of expression has by several researchers been claimed to reside in the fact that native speakers have a large number of ready-made utterances, i.e. prefabs, stored. Bolinger (1976:2) claims that "speakers do at least as much remembering as they do putting together". Similar lines of reasoning have been proposed by among others Pawley & Syder (1983), Langacker (1987), Fillmore, Kay & O'Connor (1998), Sinclair (1991), and Jackendoff (1995).

Sinclair (1991) proposed two principles that the language producer alternates between when producing language: the idiom principle and the open choice principle. These two principles incorporate the usage of complex items into language production. The open choice principle corresponds to the traditional way of looking at language production, i.e. that single lexical items are combined using a restricted set of syntactic rules, whereas the idiom principle accounts for the usage of more complex items, or in Sinclair's (ibid:110) words:

> The principle of idiom is that 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 analysable into segments. (Sinclair 1991:110)

Since the language producer alternates between the above principles, the output, i.e. the language produced, is made up of both prefabricated chunks of language and language that is produced creatively.

In order to account for the process by which the native speaker approaches the mature state described above, Wray & Perkins (2000), building heavily on Locke's theory of neurolinguistic development (Locke 1997), propose four stages with different proportions of holistic and analytic involvement in the language process. The terms holistic and analytic processing basically correspond to Sinclairs idiom and open choice principles, in that the holistic processing involves the use of ready-made chunks and the analytic processing involves creatively produced utterances. In Figure 1 below the different stages are related to the age of the speaker.

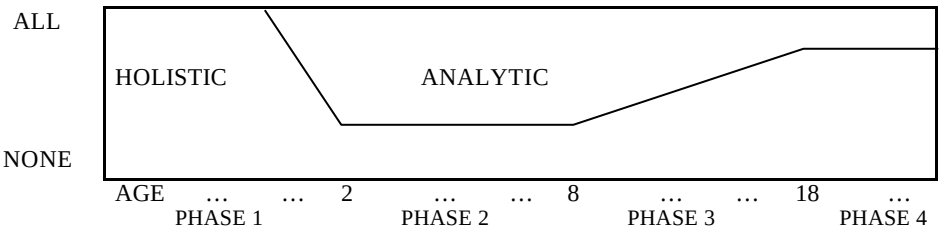

Figure 1. Relative proportions of holistic and analytical involvement in language processing from birth to adulthood (schematic representation). (taken from Wray & Perkins (2000))

In short the process can be described as follows; at first the child relies on holistic processing, i.e. in this phase the utterances are prefabricated and not analysable or put together of parts. During the later stages of that phase the cognitive capacity for analytic processing starts to take off and some utterances will be produced analytically, i.e. by means of creatively applying rules to items in the lexicon. During phase 2 the major part of all utterances are produced in this creative manner (some however remain to be produced holistically). In phase 3 the level of input, or basically the number of times each utterance has been encountered,

makes it worthwhile to start memorising certain phrases or structures (prefabs) rather than creating them anew each time they are needed. Or as Wray & Perkins (2000) put it "if the same, or similar, groups of elements are being continually encountered and/or produced it will make good economical sense to store them as separate items". The process of storing items continues into adulthood, with the result that the adult native speaker produces a major part of language holistically. It is important to note, however, that the items in the adults' holistic processing are different from the items the child uses in phase 1, something that I will return to later.

Wray & Perkins (2000) also propose that second or foreign language learners may follow a route similar to that of the native speakers. The reasons for the different phases differ, of course, between children and foreign language learners. As opposed to children foreign language learners (at least the ones who start learning the language when they are over two years of age) do not lack the mental capacity for analytic processing. Rather it is the case that they need to communicate more things than they are able to produce analytically using the grammar and words they have at their command. Often in the case of learners phase 4, the mature adult phase, is never quite reached. Going from phase 2 and 3 to phase 4 requires a massive input for the learner realise the distinction between only grammatically acceptable utterances and grammatically acceptable as well as preferred utterances. Many foreign and second language learners simply do not get enough exposure to reach that awareness.

What is also worth noticing are the differences in actual stored items in children, adults and learners. Figure 2 below is a graphic representation of what we can call the prefab stores of four different types of speakers of a language. The aphasics will not be included in the present discussion since they are not relevant to the focus of this paper. However, what we should note in this figure is that the sub-sets used by children and learners are not complete sub-sets of the adult language users' store, but rather partial sub-sets. For instance, a child may have as a stored item *Me wanna ___(slot)*, which will later disappear. Learners may have stored items that are transferred from their native language, which may disappear later in the learning process.

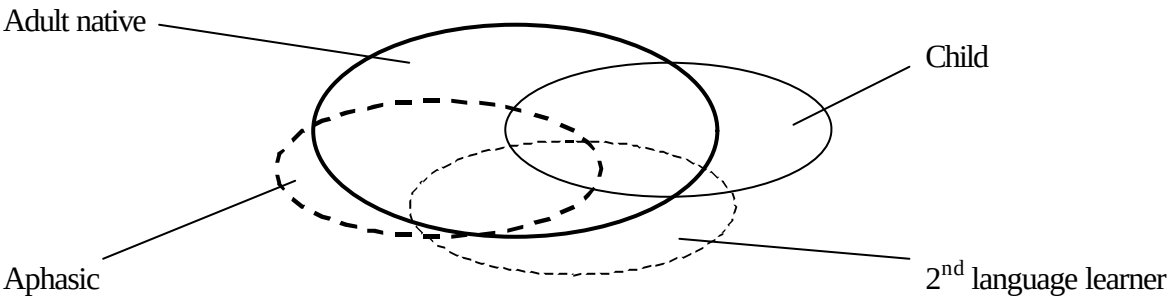

Figure 2. Formulaic expressions in different data sets as only partially coincidental (Wray 2000).

The purpose of this brief theoretical outline has been to give an insight into the lines of reasoning that lie behind the present study. I propose that a study of prefabs in learner production will give insights into the idiomatic nature of that output. I also propose that the prefab stores in both learners and natives are greatly a result of the level of exposure to the language in question. In learners, especially instructed foreign language learners, it is often the case that they never get enough exposure to achieve quite the same prefab store as the adult native speaker.

2 Previous research

In order to provide some background to the present study I will give a brief account of an earlier study (Wiktorsson 2000) on the same essays that will be further investigated here. As I said before the present study has come about as an attempt to answer some of the questions raised by that previous study.

2.1 Material and Method

The material investigated is part of the ICLE corpus, i.e. the International Corpus of Learner English. For more details concerning the corpus see Granger (1993). All learners that have contributed to the corpus are advanced learners of English, i.e. they are university students of English in their second to fourth year of study. As comparable native data the ICLE corpus also includes essays written by native English students at university level, both British and American. The material selected for Wiktorsson (2000) includes American student essays and essays written by Swedish learners. The essays are argumentative and deal with topics such as *integration or assimilation*, *environmental issues*, *modern inventions*, etc, i.e. topics that the students relate to on a rather personal level. Table 1 below gives the total number of learner and native essays used from the ICLE corpus as well as how many words they amount to.

	Number of texts	Number of words
Native	16	10907
Learner	19	10876

Table 1. Number of texts and words in the different essay categories.

Henceforth, the terms native and learner will respectively be used to represent the American students and the Swedish learners.

2.2 The prefab identification method

The essays were analysed according to a model originally created by Erman & Warren (2000). This model is designed to find the prefabs in a text and to give approximate figures for how many of the total number of words in a text are produced as parts of prefabs. The definition of prefabs that Erman and Warren work with is the following:

> a prefab is a combination of at least two words favoured by native speakers to an alternative combination which could have been equivalent had there been no conventionalization (Erman & Warren 2000)

Since this definition is based on native speaker preference, we can note that the prefabs identified will fall within the native speaker prefab store. There may of course be items in the learner essays that are prefabricated for the learner but which does not correspond to a native structure. The above definition does not cover these items.

For anything to be considered a prefab it must:

(i) consist of at least two free morphemes

(ii) manifest some feature of conventionalization, which suggests the criterion of restricted exchangeability, that is, one member of a prefab cannot be replaced by a synonymous word without causing change of meaning or function and/or idiomaticity

2.3 Types of prefabs

Figure 3 below includes the main types of prefabs identified in the essays and some of the sub-types that are found within these main types. Erman & Warren (2000), in their study, also included *reducibles*, i.e. prefabs such as *I'm, he's, didn't*, etc. but these are not included here since they are "different from the other prefabs in that they have no idiomatic meaning and no obvious functional rationale" (Warren, forthcoming).

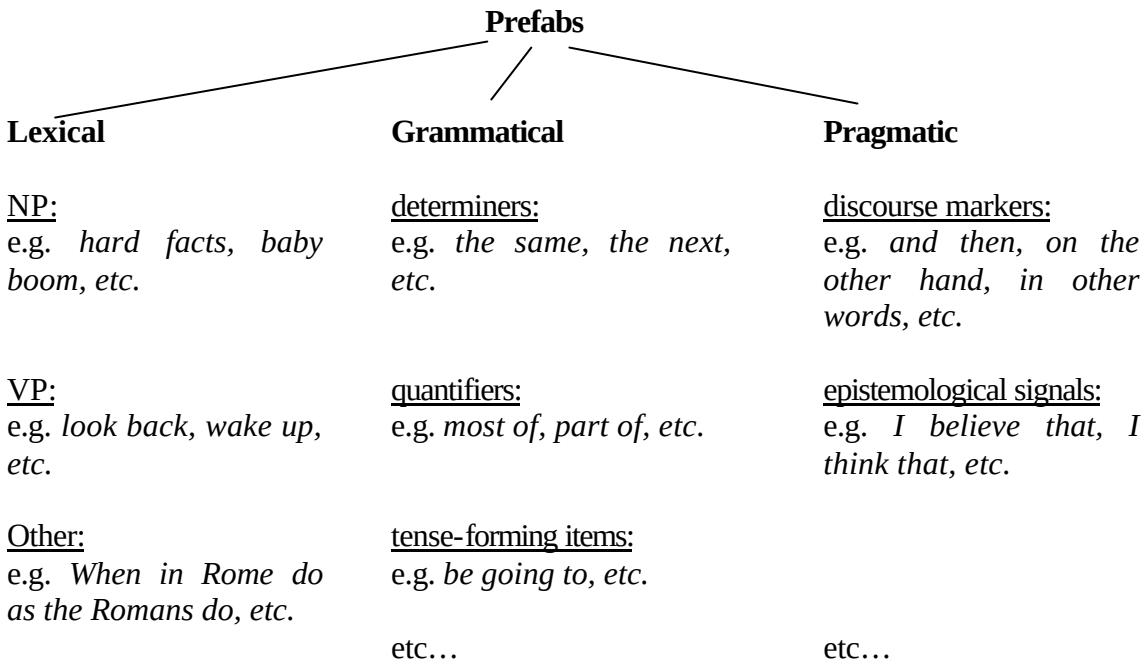

Figure 3. Main types of prefabs.

The lexical prefabs have reference and denote entities, properties, states, events, etc.. As can be seen in Figure 3 above they can be of different phrase types. The grammatical prefabs serve grammatical functions in the text, some of which can be found above. Pragmatic prefabs, of course serve pragmatic functions, e.g. as discourse markers used to indicate transitions, etc in the discourse or epistemological prefabs, which function to relieve the speaker/writer of being completely committed to the truth value of the proposition in question.

2.4 Results

Wiktorsson (2000) found that the learners use the same amount of prefabs as the native speakers, i.e. the proportion of prefabricated language in the essays investigated were the same. In Sinclair's terms the same amount of the language of both groups was produced using the idiom principle.

However, differences were found in terms of types of prefabs. The learner essays contained a higher degree of grammatical and pragmatic prefabs than the native essays,

which in turn contained a higher degree of lexical prefabs. Prefab analyses of spoken vs. written material by Wiktorsson (1998) showed similar differences in prefab type distribution between speech and writing as was found between the learner and native essays. This suggests that the learner essays are more spoken in style than the native ones. Naturally, speech differed more from writing than the learner essays do from the native essays, but since the learner essays are in fact written this is what would be expected.

Further, the lexical prefabs were analysed in more detail, which revealed a significant difference between the proportions of verb phrases and noun phrases. Of the total lexical phrases in the learner essays 46,7% were VP's and 37,2% NP's and in the native phrases 34,4% were VP's and 55,2% NP's. Again this difference has been found between speech and writing, with writing then containing a higher degree of nouns due to higher lexical density and more nominalizations (De Vito 1964) (Chafe & Danielewicz 1987).

The pragmatic prefabs found were also analysed further and the sub-types compared. The learner essays were found to contain a higher degree of pragmatic prefabs that would typically be found in speech, such as interactives (e.g. *So, what do we do?* and *you see*).

3 This study

This study will take off where the previous one left off. The lexical and pragmatic prefabs found by Wiktorsson (2000) will be further investigated in order to see if the learners actually use prefabs that are more spoken in nature. This will be done be checking the prefabs in two corpora, one of speech and one of writing. Since the pragmatic prefabs are rather limited in number all of these will be investigated. The lexical prefabs, however, are too many to work with, and therefore a representative selection is sampled from them in order to get a manageable number to work with.

4 Material and Method

4.1 The sub-set of prefabs selected for the present study

It was not possible to analyse all the prefabs found by Wiktorsson (2000), because the corpus work would have been unmanageable. The first exclusion that was made was to focus only on pragmatic and lexical prefabs since these prefabs will reveal more about the style value of the texts than the grammatical prefabs will do. Secondly, all repeated prefabs were excluded in order to get at the prefab store, i.e. what different types of prefabs were used rather than all the different tokens of these. This gave the following number of prefabs:

	Native	Learner
Lexical prefabs	755	813
Pragmatic prefabs	74	102

Table 2. Number of lexical and pragmatic prefab types in the native and learner essays.

These pragmatic and lexical prefabs then constitute the different types that were found in the native and learner essays. Some of these prefab types were found in both varieties, as can be seen in Table 3 below.

	Native			Learner			Common
	All	Variety specific	%	All	Variety specific	%	Common to both varieties
Lexical	755	699	92,6%	813	757	93,1%	56
Pragmatic	74	59	79,7%	102	87	85,3%	15

Table 3. Lexical and pragmatic prefabs divided into specific and common ones.

Most of the lexical prefabs in each variety were only found in that variety, 92, 6% for the native ones and 93,1% for the learner ones. The percentages of variety specific were slightly lower for the pragmatic prefabs (79,7% for the native and 85,3% for the learner), which would be expected since pragmatic prefabs constitute a more restricted set.

In Table 3 above there are six different sets of prefabs. The pragmatic prefabs are few enough for all to be included in the study, but for in the lexical prefab categories only noun phrases and verb phrases will be included. The reason for this exclusion is that the NP's and VP's constitute the absolute majority of the lexical prefabs; 85% of the native lexical prefabs are NP's or VP's, 83% of the learner ones and 82% of the common ones. This resulted in the following numbers of prefabs left in the different categories.

	Native	Learner	Common
VP's	268	355	25
NP's	323	272	20

Table 4. Number of NP's and VP's in the three different lexical categories.

Since there were so few prefabs in the common category it was possible to keep that intact. The specific NP and VP categories still contained more prefabs than it was possible to work with. Therefore a representative selection of approximately 100 prefabs from each type was sampled, i.e. a total of around 400 prefabs.

The resulting number of prefabs used for this study can be found in Table 5 below.

Selection	Type	Number
All	Pragmatic common	16
All	NP common	20
All	VP common	25
All	Native specific pragmatic	59
All	Learner specific pragmatic	89
Sample	Native specific NP	~100
Sample	Learner specific NP	~100
Sample	Native specific VP	~100
Sample	Learner specific VP	~100

Table 5. Number of prefabs in each category used for testing register differences.

4.2 The corpus tests

The Bank of English corpus (see http://titania.cobuild.collins.co.uk/boe_info.html) was used for testing whether the prefabs selected were more frequent in speech or in writing. From the corpus the following sub-corpora were selected:

Sub-corpus	Total number of words
Spoken	9272579
Written	8888115

Table 6. The number of words in the sub-corpora selected from The Bank of English corpus.

As we can see there is a slight difference in the number of words in these two sub-corpora (384464 words). However, since speech and writing differ very much in terms of lexical density the actual number of instances of each prefab in each of the two corpora will be irrelevant and so will consequently the number of words in each sub-corpus. What will be compared are simply the "relative differences", i.e. how many times more common the prefabs in the different sets are in speech vs. writing.

All the individual prefabs in the different lists were checked in the spoken and written corpora and the numbers of instances found were counted. However, approximately 10 of the VP's in each specific list proved impossible to check since the search engine used did not support the search for discontinuous items.

5 Results

Wiktorsson (2000) found qualitative differences that might be related to a more spoken style in the learners' output. In the present study the evidence partly support this hypothesis.

In Table 7 below we see the results from the frequency checks in the spoken vs. written corpora. The pragmatic prefabs that were found in both categories were more frequent in speech than writing, but only 2,6 times more. These are all fairly general and unmarked ones such as: *and so on, first of all* and *on the other hand*.

Prefab category	#	Written	Spoken	Times	More common in
Pragmatic common	16	4426	11876	2,68	Speech
Native specific pragmatic	59	5610	34547	6,16	Speech
Learner specific pragmatic	89	4151	74742	18,01	Speech
NP common	20	1950	1370	1,42	Writing
Native specific NP	103	6835	4372	1,56	Writing
Learner specific NP	109	6629	3175	2,09	Writing
VP common	25	6876	6237	1,10	Writing
Native specific VP	90	7632	5251	1,45	Writing
Learner specific VP	91	11136	16929	1,52	Speech

Table 7. Comparison of frequencies of occurrences in speech vs. writing for the different prefab categories.

If we compare the ones found only in each of the two text-types, we find that the learner essays contain pragmatic prefabs that are very much more frequent in speech than the native

ones. The pragmatic prefabs found in the native essays occur 6,16 times more often in the spoken corpus whereas those found in the learner essays occur 18,01 times more often in speech. This clearly tells us that the learners do use pragmatic prefabs that are of a much more spoken nature than the learners do.

The NP's in all categories are more common in writing, which is not surprising because of the lexical density in writing is higher and so is the ratio of nouns. There is in principle no difference between the common ones and the native specific ones, the former are 1,42 times more common and the latter 1,56 times more common. The learners' NP's are slightly more common, 2,09 times. This might argue against learners using prefabs that are more spoken in style, at least when it comes to NP's.

The learner specific VP's are approximately as much more common in speech as the native ones are more common in writing. As opposed to the findings for the previous category these results support the hypothesis that the learners use prefabs of a more spoken kind. What can also be noted about the learner VP's is that they are a lot more common both in speech and writing than the native or the common ones. This suggests that these are high frequency items that we can assume that the learners have come across several times and thus the level of input where items of this type become stored has been reached.

6 Conclusion

For two of the three types of prefabs investigated, VP's and pragmatic prefabs, we have found that the learners use prefab items that are more frequent in speech then the natives do. In Wiktorsson (2000) we also found that the learners used more VP's and pragmatic prefabs than the natives. Both these findings indicate a more spoken style in the learners' output.

Let us return to the different prefab stores of natives and learners that were indicated by Wray's figure (Figure 2) in section 1.1. Something that is relevant to the present study is that there within the native circle must exist sub-circles that are related to register. Some prefabs are found more in written production and some more in spoken production. These circles of course overlap, since most items are possible to use in both modes of production, but with differences in probability of occurrence.

In this study we have focused on the part of the learner prefab store that falls within the native store. Regardless of that we can assume that the fuller picture is as indicated by Figure 4 below. I.e. the Swedish learners' prefab store includes items that have a higher likelihood of being used by natives in speech than in writing.

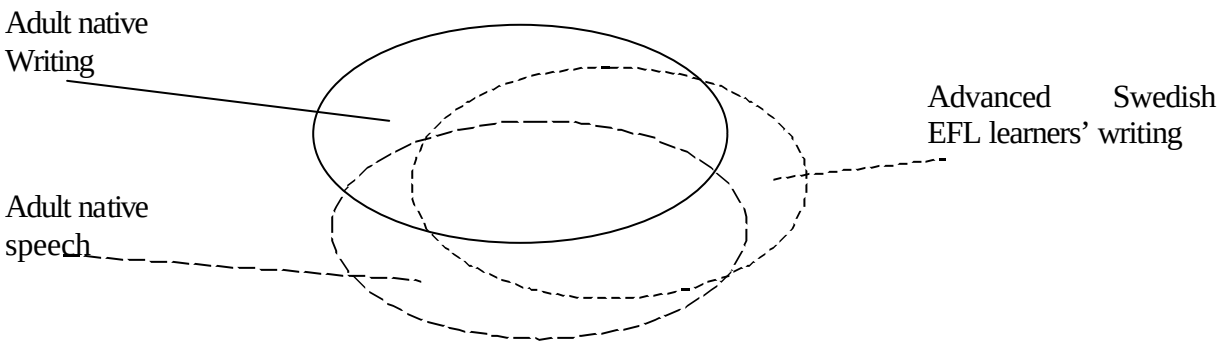

Figure 4. Possible model of learner prefabs as compared to native prefabs.

Prefabs need to have been encountered many times before they become stored. This importance of input should be taken into consideration here, and one source of input that cannot be disregarded is that of television. The majority of TV-shows in Sweden are British or American in origin. Since these are not dubbed it is fair to assume that many young people today get a major part of their English input through television. Even though the dialogues etc. in TV-shows are originally written, they are written to be spoken. Thus a major part of our learners input can be claimed to be spoken in nature. Perhaps this is what we see reflected in their output, which is not unidiomatic per se, but rather too spoken in style.

7 References

Bolinger, D. L. 1976. Meaning and Memory. *Forum Linguisticum* 1: 1-14.

Chafe, W. & J. Danielewicz. 1987. Properties of Spoken and Written Language. In *Comprehending Oral and Written Language*, ed. Horowitz, R. & S. J. Samuels, San Diego: Academic Press.

Erman, B. & B. Warren. 2000. The Idiom Principle and the Open Choice Principle. *Text* 20: 29-62.

Fillmore, C. J., P. Kay & M. C. O'Connor. 1998. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64: 501-538.

Granger, S. 1993. The International Corpus of Learner English. In *English Language Corpora: Design, Analysis and Exploitation*, ed. Aarts, J., P. de Haan & N. Oostdijk, Amsterdam: Rodopi.

Jackendoff, R. 1995. The Boundaries of the Lexicon. In *Idioms: Structural and Psychological Perspectives*, ed. Everaert, M. e. a., Hillsdale, New Jersey: Lawrence Erlbaum.

Langacker, R. W. 1987. *Foundations of cognitive grammar, Volume 1, Theoretical Prerequisites*. Stanford, Calif.: Stanford University Press.

Locke, J. L. 1997. A Theory of Neurolinguistic Development. *Brain and Language* 58: 265-326.

Pawley, A. & F. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In *Language and communication*, ed. Richards, J. C. & R. W. Schmidt, London, New York: Longman.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford, New York: Oxford University Press.

Warren, B. forthcoming. An Alternative View of Stored Linguistic Knowledge and its Relevance to Text Composition. *Text*

Wiktorsson, M. 1998. Compositional and Non-Compositional Aspects of Written and Spoken Texts. Paper presented at *CSDL4*, October 10-12, 1998, Atlanta, USA.

Wiktorsson, M. 2000. Prefabricated phrases in learner language. A corpus-based study comparing advanced EFL writing with native English writing. In *Korpusar i forskning och undervisning*, ed. Byrman, G., H. Lindquist & M. Levin, Uppsala:

Wray, A. 2000. The functions of formulaic language. Paper presented at *AAAL 2000*, 29 February 2000, Vancouver.

Wray, A. & M. R. Perkins. 2000. The functions of formulaic language: an integrated model. *Language and Communication* 20: 1-28.

Yorio, C. A. 1989. Idiomaticity as an indicator of second language proficiency. In *Bilingualism across the lifespan*, ed. Hyltenstam, K. & L. K. Obler, Cambridge: Cambridge University Press.