JOURNAL OF MEMORY AND LANGUAGE 33, 357–375 (1994)

Prefix Stripping Re-Revisited

ROBERT SCHREUDER

Interfaculty Research Unit for Language and Speech, Nijmegen University, The Netherlands

AND

R. HARALD BAAYEN

Max-Planck-Institut für Psycholinguistik, Nijmegen, The Netherlands

Taft and Forster (1975) enriched serial search models of lexical access with a prefix-stripping module, hypothesizing that prefixal morphology is functional in the sense that it leads to more efficient lexical processing. A detailed investigation of the lexical statistics of prefixation and pseudo-prefixation in English and Dutch shows, however, that the addition of a prefix stripping module to a serial search model leads to a substantial decrease in its processing efficiency, defined as the number of search steps required for accessing words in the mental lexicon. This reversal is attributed to the high token frequencies with which pseudo-prefixed words occur. We also briefly discuss and compare the lexical statistics of a similar model in which obligatory segmentation takes place, the metrical segmentation theory proposed by Cutler and Norris (1988). Methodological implications for the testing of psychological models of language processing are discussed.

Many studies have been carried out that have investigated the role of morphological complexity in word recognition. The basic question these studies try to answer is whether access to the lexicon is influenced by the fact that words may have differing internal morphological structure. Do the constituents of a morphologically complex word play a role in lexical access? This question has led Taft and his co-workers to propose an influential model, the "prefix stripping" model (Taft & Forster, 1975, 1976; Taft, 1979a, 1979b, 1981, 1985; Taft, Hambly, & Kinoshita, 1986; Bradley & Forster, 1987; Taft, 1988). This model has inspired numerous further investigations (e.g., Bergman, Hudson, & Eling, 1988; Colé, Beauvillain, & Ségui, 1989; Burani & Caramazza, 1987; Lima, 1987; Manelis & Tharpe, 1977; Schriefers, Zwitserlood, & Roelofs, 1991) and it is the topic of the present article.

In this article we present a critical evaluation of the prefix stripping model. We will do so not by discussing the many experimental findings (for a good review see for instance Henderson, 1985) but by focusing in particular on the computational aspects of the model. We will do so by presenting data on lexical statistics, employing large corpora of texts in English and Dutch. We will then present some computational data that evaluate the results of our lexical statistics in terms of processing efficiency. We will argue that these data show that prefix stripping does not enhance processing efficiency for serial search models as originally claimed by Taft and Forster, at least for languages like English and Dutch. A more general aim is to show that empirical investigation other than experimentation

Both authors contributed equally to the article. They are indebted to Maarten van Casteren, Ton Dijkstra, Henk van Jaarsveld, Marcus Taft, and an anonymous referee for valuable discussion. Reprint requests should be addressed to R. Schreuder, Interfaculty Research Unit for Language and Speech, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. (Electronic mail: rob@mpi.nl)

Accepted by Dr. Shoben.

0749-596X/94 $6.00

may guide the evaluation of theoretical models. This is illustrated by a comparison with the theory of metrical segmentation proposed by Cutler and Norris (1988), a theory that also predicts obligatory stripping of word initial segments.

The Prefix-Stripping Model

The prefix stripping model is a particular instantiation of a linear search model of lexical access. Following the outline of Taft (1988), the model can be described as subdividing the lexicon in two components: an input system (actually two, one for the visual modality and one for the auditory modality) and a (modality free) central system. After having achieved a satisfactory match between a presented word and a representation in the frequency ordered input system, information about that word is made accessible from the central system by means of pointers from the input system to the central system. This information is assumed to concern the meaning of a word, its syntax, as well as spelling and pronunciation information. The basic assumption of the model is that the representations of derivationally complex words in the input system are representations of their stems. The representations in the central lexicon are of the complete word.

By way of example, consider the recognition of the morphologically complex word *reforest* in either the visual or the auditory modality. After determination of which access system (or "access bin") is relevant, the sensory to lexical match is made on the basis of the stem *forest* or alternatively the so-called BOSS (basic orthographic syllabic structure, Taft, 1979b) in the visual modality and on the basis of the stem in the auditory modality. In order for this match to be made, the presented letter string or utterance must be stripped of its prefix. The resulting stem is then used to locate a matching entry by means of a serial search in the (relevant) access bin, the entries of which are assumed to be ordered on the basis of the cumulated frequencies of the words they occur in. The matching entry provides a pointer to the central system, that is, to the location of all words containing that particular stem (or BOSS). Since Taft (1979a) claims that whole word frequency co-determines search times, we assume that the list (or bin) of word representations in the central system for this particular stem (or BOSS) is also ordered by frequency, but in this case on the basis of the frequency of the whole word.[1] Hence the central system is assumed to contain bins with all morphologically related stems (in this case, e.g., the stem entry *forest* and its subentries *reforest* and *deforest*), listed in order of decreasing frequency. This bin then is searched serially until a match with the original input is found.[2] The frequency ordering in both systems can explain frequency effects obtained in word recognition studies both for frequency of stems and frequency of full forms (see for instance Taft, 1979a).

The origins of serial search models can be traced to programming techniques, where through so-called "hash" coding the region where the relevant information is stored is determined. Serial search subsequently looks for the precise location in this region (bin). In Taft and Forster's serial search model, the search key is the initial constituent of the word, that is, the BOSS or the first morpheme. What considerations motivate the exceptional status of prefixed words, where the search key is obtained after stripping away the first morphological constituent? The main theoretical motiva-

[1] Although this point has not been made explicit by, e.g., Taft (1979a), some mechanism for selecting the appropriate word representation from the (often substantial) set of words pointed to by an access representation has to be found. Within the present framework the most natural solution is to assume that candidate selection proceeds on the basis of a serial search through the frequency-ordered list of the relevant word representations.

[2] Incidentally, this means that besides having the stem representation, the central system needs a copy of the full input representation in order to compare it with the entries in the central system. How this can be achieved with a central system that is supposed to be modality free is never explained.

tion appeals to a supposed increase in the processing efficiency of the serial search (Taft & Forster, 1975, pp. 645–646):

> . . . Knuth (1973) has suggested that by stripping off the prefix, *re-* for example, one can use an alphabetical storage of words without having to list a very large number of words under the same description. Thus it is possible that a system using morphological decomposition would access prefixed words faster than a system which left the word intact, since the entry for *rejuvenate* could be located without having to search through all the words beginning with *re-*.

In the following sections the claim that the implementation of prefix stripping in a serial search model enhances its processing efficiency will be shown to be incorrect. We will do this by evaluating the feasibility of the prefix stripping design, focusing on the statistical properties of the lexicons of English and Dutch as a complementary methodology to experimental approaches to testing the prefix stripping hypothesis.[3]

The first question we will deal with concerns the frequency with which the parsing system has to process pseudo-prefixed words such as *indexing* and *decorum*. Since the pseudo-stem cannot be located in the access bin, the system has to initialize a second search on the basis of the full form (backtracking). We will investigate how often lexical look-up is slowed down in this way due to such backtracking. Subsequently, we will evaluate the computational costs of a prefix stripping retrieval system.

Two methodological choices underlie the research reported here. First, we have opted for giving the prefix stripping theory the benefit of the doubt. That is, whenever forced to make implicit assumptions of the model explicit, we will be as lenient as possible to the prefix stripping theory by selecting those solutions that render it more effective. Second, since the present paper focuses on the prefix stripping hypothesis as such rather than on the validity of serial search models in general, we will adopt a serial search model without a prefix stripping component as the base line condition. This will allow us to evaluate Taft and Forster's (1975) claim that the addition of a prefix stripping module to a serial search model will lead to an increase in processing efficiency.

[3] For a similar approach to problems of lexical density see Frauenfelder, Baayen, Hellwig, & Schreuder, 1993.

The Statistics of Backtracking in a Serial Search Model

This section presents the lexical statistics of backtracking in a serial search model. Backtracking occurs in a number of circumstances. Pseudo-prefixed words are the predominant cause of backtracking (e.g., *reindeer*). Another type of backtracking is triggered by words such as *begin*, where *gin* is an existing stem that foils efficient retrieval of *begin* itself. Similarly, allomorphy and vocalic alternation may cause backtracking too, even though a true prefix is involved.

Following a discussion of these different cases of backtracking we present frequency counts for a series of English and Dutch prefixes and pseudo-prefixes, in terms of both their orthographical and their phonological representations. The counts are based on the CELEX lexical databases of English and Dutch (Burnage, 1990, Baayen, Piepenbrock, & van Rijn, 1993), the frequency counts of the English database pertaining to the frequencies of occurrence in the Cobuild corpus,[4] and those of the Dutch database to the frequencies of occurrence in the INL corpus.[5]

Types of Backtracking

Backtracking caused by pseudo-prefixation. Since the lexical statistics of backtracking to be presented below depend largely on what counts as a pseudo-prefixed word, we first determined some reasonable criteria for pseudo-prefixation. Let us as-

[4] See e.g. Renouf (1987). The Cobuild corpus is taken from both written (75%) and spoken (25%) language, covering a wide variety of forms of English.

[5] The INL corpus contains roughly 42 million word tokens of written texts. It has been compiled by the Dutch national institute for lexicology.

sume as a working definition that a pseudo-prefixed word is characterized by the property that it contains a prefixal string followed by a non-occurring stem. The properties of the non-occurring stem are crucial here. Two sets of properties are relevant. The first set concerns form properties of the non-occurring stem. The second set concerns the status of bound stems.

First, consider the processing of word types such as *retch* and *renderings*. Having stripped the pseudo-prefix *re*, one is left with a pseudo-stem that cannot be syllabified. If one is willing to assume that the phonotactics of the stripped stem are evaluated before the serial search is initiated, these stems should not be counted as true pseudo-stems. Hence we have not counted words such as *retch* and *renderings* as pseudo-prefixed words, thus ensuring that all our pseudo-stems are phonologically and orthographically legal. Thus our first criterion for a legitimate modality-stem is that

(1) A pseudo-stem should be syllabifiable.

This criterion applies to both modalities.

Another related issue concerns the length of a potential pseudo-stem. Consider words like *beach* and *reach*. In contrast to *react*, where decomposition is possible, and *retch*, where syllabification fails after stripping, the putative pseudo-stem *-ach* might be argued to be too short to count as a possible stem. More precisely, if the putative stem is so short as to have a very low probability of existing in the language, we can enrich the prefix stripper with the strategy of stripping only relatively long words. Since this strategy requires the early availability of information on word length, we have used this strategy for the visual modality only. The criterion for discounting a certain word length has been that for such a word length not more than 10% of all word tokens should contain truly prefixed word tokens. That is, prefix stripping is not carried out for shorter words as long as this strategy does not lead to an error rate larger than 10% for those words that should in fact be parsed. In short:

(2) Relatively short orthographic strings are not stripped.

These choices eliminate a substantial number of potential pseudo-prefixed words from consideration. Note that these criteria presuppose the availability of linguistic information at very early processing stages, which as such may well be questionable. Our choice not to include the (many) words with these "trivial" pseudo-stems in our counts is based on the methodological decision of giving the prefix stripping hypothesis the benefit of the doubt.

Second, consider the phenomenon of bound stems, stems that occur in combination with some affix, but which do not constitute independent lexical items. The problem here is that we have to establish criteria for distinguishing between bound stems (*-ject* < *re-ject*) and pseudo-stems (*-corum* < *decorum*). A first criterion concerns the semantic transparency of the pattern:

(3) A putative pseudo-stem is a bound stem in case it occurs in at least one semantically fully transparent combination.

Following Aronoff (1976), we assume that words such as *refer, defer, prefer, repel* or *remit, permit,* and *admit* are not semantically transparent. Hence, we analyze them as containing pseudo-stems. Consequently, our counts take prefix stripping to lead to processing errors for words like this. In contrast, as argued by Corbin (1985), verbs like *ingress, egress,* and *progress* may well be semantically transparent. These verbs can indeed be argued to contain bound stems. Hence prefix stripping does not lead to processing errors for such words.

This criterion of semantic transparency is required by psycholinguistic considerations. The only way in which the language learner can discover that a certain string actually is a stem (free or bound) is when that string occurs in at least one semantically fully transparent combination. In the absence of any semantic transparency, the

language learner will not consider the string *ar* to be a stem, even though it is found in strings containing pseudo-prefixes such as *be-ar, re-ar, de-ar,* and *e-ar.*

A second, closely related criterion concerns the possibility of bound stems to participate in productive word formation, as pointed out by Corbin (1985). Hence:

(4)A putative pseudo-stem is a bound stem in case it participates in productive word formation.

In fact, participation in productive word formation provides supporting evidence for the transparency of the bound stem. For example, the Dutch verbs *ontginnen* ("bring into cultivation") and *beginnen* ("to begin") contain the prefixes *be-* and *ont-* attached to the synchronically pseudo-stem *gin,* historically a verb that has been lost in the present-day language. Although the inflectional pattern still reveals its complex origin—monomorphemic verbs such as *lopen,* "to walk", take the prefix *ge-* for the past participle (*gelopen*), whereas prefixed words such as *ontlopen,* "to run away", and *belopen,* "to walk on", do not; here, the past participles are again *ontlopen* and *belopen,* and the same applies to *beginnen* and *ontginnen*—the pseudo-stem *gin* has no meaning of its own and does not participate in any word formation. Hence words like *beginnen* and *ontginnen* are counted as cases where prefix stripping leads to processing errors, the stem having no entry in the access bin.

We have used criteria (3) and (4) in the lexical counts to be presented below.[6] In the case of the English counts the morphological analyses of CELEX have been used, which are based on *The Shorter Oxford English Dictionary* (1973) and the *Collins Dictionary of the English Language* (1986). Both dictionaries apply our criteria, with the proviso that lexicographic practice tends to be conservative at times, notably with respect to bound stems like *-gress.* As in Aronoff (1976), *-gress* (according to our criteria a bound stem) is actually handled in the same way as *-mit* and *-ject,* that is, as a pseudo-stem. The analyses of Dutch prefixes is based on the CELEX parsings, supplemented by the analyses in the careful study by de Vries (1975) of verb formation in Dutch.

We believe that the definition of pseudo-stems developed here is better motivated than the rather vague notion of pseudo-affixation found in Taft and Forster (1975, 1976), both from a psycholinguistic and a linguistic point of view. However, in order to ascertain whether one's definition of pseudo-prefixation plays a major role, we have also carried out some analyses which follow as close as possible the original descriptions.

Other cases of backtracking. Prefix stripping runs into problems not only in the case of pseudo-prefixed words but also in a variety of other cases where a real prefix is involved. Consider the following Dutch prefixed words:

(a) *ver-bind*	"to bandage"	*ver-band*	"bandage"
(b) *ver-lies*	"to lose"	*ver-loor*	"to lose", past tense
(c) *be-wust*	"conscious"	*(weten)*	"to know"

After stripping of the prefix in the noun *verband* (a) one is left with the stem *band,* "tire", "link", or "band". It is unlikely that *verband,* "bandage", is stored with *band.* If anything, it is to be located in bin of *binden,* "to bind", where *verbinden,* the verb to which it is most closely related semantically, is to be found. Hence prefix stripping leads to an unexpected kind of backtracking. The access lexicon has a pointer to *band,* nevertheless, the master lexicon does not list *verband* there. Hence one has to re-initiate the access search. Since *band* does not match with *bind,* it must be assumed that *verband* has its own

[6] Note that Taft & Forster's criteria for pseudo-prefixation are quite vague. For instance, Taft & Forster (1975; p. 644) assume that *bezzle* is a stem of English where prefix stripping is applied successfully. Although *bezzle* is listed as a free stem in Webster (1981), we doubt that their experimental subjects were familiar with this lexicographic oddity. More importantly, the resulting lack of semantic transparency for *embezzle* implies that *embezzle* should be treated as a pseudo-prefixed word.

access entry, even though it is morphologically complex and semi-transparent. Similar problems arise in English with words such as *begin*.

In the verb *verlies* (b) the prefix attaches to a verbal base that is similar to *-gin* in *begin* and *ontgin* in that it never appears as a verb in isolation. In this case, however, the verbal base has the noun *lies*, "groin", as a homophone. With *lies* as search key for the verb *verlies*, a search in the access bin will be successful. For *verlies*, however, the access bin provides a pointer to an entry in the central storage system where the noun *lies* and its inflected variants, its compounds and its derivations are listed, but not the verb *verlies* itself. As in the case of *verband*, the result is an erroneous search in the central lexicon. To complicate matters even further, the past tense *verloor* contains the otherwise unattested string *loor*, that can hardly be assumed to have independent existence. Evidently, *verloor* must be assumed to have its own access entry. The last example, (c) concerns a stem that is historically related to the verb *weten*, "to know", but that appears only in the one adjective *bewust*. Although one could build one's model such that the access entry *-wust* points to the central bin of *weten*, there is little linguistic support for such a move, apart from the fact that it is unclear how the language learner would acquire such a lexical organization.

Having discussed the criteria defining pseudo-stems and having illustrated some of the difficulties that arise when implementation of prefix stripping is considered in detail, we now turn to the lexical statistics of prefix stripping.

Token Counts

Lexical processing is token-based rather than type-based. Word frequency distributions are highly skewed. Generally, roughly half of the different words (lemmas, or types) in such counts occur once only. Conversely, a relatively small number of different word types account for a large portion of all word tokens. Thus we find that the monomorphemic words in the Dutch corpus underlying our frequency data, a corpus of roughly 40 million word tokens, account for 79% of the tokens but only for 12% of the types (see Chitashvili & Baayen, 1993). Since word types appear with such widely varying frequencies in language, one cannot evaluate models of lexical access on the basis of a vocabulary list. Each word in such a list has to be weighted for the frequency with which it occurs in some corpus. One cannot focus on the way individual types are processed *independently* of the frequencies with which these types appear in the language. In the case at hand, we therefore present the token frequencies with which truly prefixed words and pseudo-prefixed words appear in Dutch and English text corpora.

METHOD

Materials

The token counts to be presented below are based on the frequencies of occurrence of all word types that appear in the Dutch INL corpus and the English Cobuild corpus as available in the CELEX lexical database.

Procedure

The most frequently occurring prefixes of English and Dutch were selected for further investigation. The Dutch verb-forming prefixes *be-*, *ver-*, *ont-*, and *her-* as well as the adjective forming prefix *on-* (English *un-*) were analyzed. For English we studied the prefixes *be-*, *de-*, *en-*, *re-*, *un-*, *in-*, *mis-* (see Baayen & Lieber, 1991 for a detailed discussion of the linguistic and frequential properties of these prefixes). We then queried the CELEX databases for all word types and their frequencies containing an initial string matching these prefixes. Two such queries were conducted, one for orthographic word representations, the other for phonological word representations. Note that there may be crucial differences between the two representations of a single

dictionary entry. For instance, the prefix will be correctly analyzed as such in both the orthographic and the phonological representation of a word like *reforest*. In contrast, a word like *regatta* will be analyzed as containing a pseudo-prefix in the orthographical representation but not in the phonological representation.

We classified these word types as follows. In the case of English we made a distinction between three categories: fully regular prefixed words (*reforest*), semantically opaque prefixed words (*reformation*), and pseudo-prefixed words (*reindeer*). The criteria for pseudo-prefixation have been discussed above. The criterion used by CELEX for non-transparency is the absence of even a single semantically transparent reading. In the case of Dutch, we carried through a somewhat more fine-grained classification, using our intuitions as native speakers of Dutch and building on the detailed analyses of verbal prefixation in Dutch by de Vries (1975). In addition to fully regular words (REG) and pseudo-prefixed words (PSE) we distinguished between the following cases: stem-allomorphy and vocalic alternations as in *verband* (bandage) (ALM), prefixed words with opaque bound stems that are formally complex only (FCO) (*verliezen*, to lose), formations with doubtful transparency (*verraden*, to betray) (DTR), and completely non-compositional, opaque words (*verrichten*, to carry out) (OPQ). The analysis was agreed upon by three native speakers of Dutch. Having made these classifications we can now establish the number of tokens within each category. This somewhat more fine-grained classification will allow us to assess how often truly prefixed words that are fully transparent are in fact encountered, as well as to gauge the influence of allomorphy and vocalic alternation on the efficiency of the access algorithm.

Results and Discussion

Tables 1 and 2 list the results obtained for the selected Dutch and English prefixes. As can be seen in Table 1, the percentage of pseudo-prefixed orthographical word tokens ranges from 2% in the case of *ver-* to 59% in the case of *on-*. Phonological word tokens show a similar pattern. To evaluate the error rate of prefix stripping, we will count as an error those cases involving pseudo-prefixation, allomorphy or vocalic alternation. Across all Dutch prefixes studied here this leads to an overall error rate of 30.4% for orthographical wordforms and 29.4% for phonological wordforms.

Turning to English (Table 2), we observe that the percentage of pseudo-prefixed orthographical word tokens ranges from 24% in the case of *un-* to 98% in the case of *de-*. For phonological wordforms this percentage ranges from 44% for *un-* to again 98% for *de-*. Across all English prefixes studied here this amounts to an overall error rate of 81% for orthographical wordforms and 83% for phonological wordforms.

The present results strongly suggest that incorporating a prefix stripping mechanism in a serial search model leads to unexpectedly high error rates for languages such as Dutch and English. The simple fact that prefix stripping gives rise to so many errors and hence requires many backtracking operations is prima facie evidence against the incorporation of prefix stripping in a serial search model.

However, it might be possible that despite these high error rates a serial search algorithm incorporating prefix stripping is nevertheless computationally more efficient than a serial search without a prefix stripping module: the costs of backtracking may turn out to be less than the costs involved when linear searches have to be carried out in central bins, each of these bins containing all words with a particular (real) prefix. This issue is the topic of the next section.

Processing Efficiency

We evaluate the computational efficiency of the prefix stripping hypothesis by comparing the number of search steps required

TABLE 1
TYPE AND TOKEN COUNTS FOR PREFIXES AND PSEUDO-PREFIXES FOR DUTCH ON THE BASIS OF THE INL CORPUS AS AVAILABLE UNDER CELEX

		Orthography				Phonology			
Prefix	Category	Types	(%)	Tokens	(%)	Types	(%)	Tokens	(%)
be-	REG	918	(39.67)	229,218	(30.98)	938	(45.76)	28,862	(31.58)
	DTR	272	(11.75)	113,038	(15.28)	296	(14.44)	123,050	(16.98)
	OPQ	475	(20.53)	217,625	(29.41)	492	(24.00)	219,130	(30.24)
	FCO	177	(7.65)	87,052	(11.77)	185	(9.02)	87,052	(12.01)
	ALM	62	(2.68)	11,226	(1.52)	56	(2.73)	9,792	(1.35)
	PSE	410	(17.72)	81,767	(11.05)	83	(4.05)	56,728	(7.83)
	T	2314	(100.00)	739,916	(100.00)	2050	(100.00)	724,614	(100.00)
ver-	REG	1438	(68.48)	346,774	(56.65)	1438	(69.54)	346,774	(56.50)
	DTR	155	(7.38)	75,364	(12.31)	155	(7.50)	75,364	(12.28)
	OPQ	153	(7.29)	46,089	(7.53)	182	(8.80)	54,272	(8.84)
	FCO	194	(9.24)	99,479	(16.25)	194	(9.38)	99,479	(16.21)
	ALM	82	(3.90)	31,005	(5.07)	54	(2.61)	22,852	(3.72)
	PSE	78	(3.71)	13,394	(2.19)	45	(2.18)	15,043	(2.45)
	T	2100	(100.00)	739,916	(100.00)	2068	(100.00)	613,784	(100.00)
her-	REG	144	(42.48)	12,047	(27.26)	145	(43.81)	12,053	(28.60)
	DTR	17	(5.01)	2,306	(5.22)	17	(5.14)	2,306	(5.47)
	OPQ	27	(7.96)	9,623	(21.77)	27	(8.16)	9,623	(22.82)
	FCO	0	(.00)	0	(.00)	0	(.00)	0	(.00)
	ALM	5	(1.47)	11,811	(26.73)	7	(2.11)	11,971	(28.39)
	PSE	146	(43.07)	8,407	(19.02)	135	(40.79)	6,220	(14.75)
	T	339	(100.00)	44,194	(100.00)	331	(100.00)	42,173	(100.00)
on-	REG	1121	(54.44)	112,324	(34.12)	944	(52.13)	98,828	(30.29)
	DTR	42	(2.04)	3,469	(1.05)	38	(2.10)	3,454	(1.06)
	OPQ	23	(1.12)	1,510	(.46)	24	(1.33)	1,578	(.48)
	FCO	26	(1.26)	8,327	(2.53)	16	(.88)	7,852	(2.41)
	ALM	8	(.39)	8,261	(2.51)	8	(.44)	8,261	(2.53)
	PSE	839	(40.75)	195,342	(59.33)	781	(43.13)	206,286	(63.23)
	T	2059	(100.00)	329,233	(100.00)	1811	(100.00)	326,259	(100.00)
ont-	REG	207	(42.51)	28,558	(22.40)	207	(42.33)	28,558	(22.25)
	DTR	33	(6.78)	6,351	(4.98)	33	(6.75)	6,351	(4.95)
	OPQ	182	(37.37)	85,382	(66.99)	183	(37.42)	86,201	(67.18)
	FCO	23	(4.72)	3,495	(2.74)	23	(4.70)	3,495	(2.72)
	ALM	2	(.41)	128	(.10)	2	(.41)	128	(.10)
	PSE	40	(8.21)	3,549	(2.78)	41	(8.38)	3,589	(2.80)
	T	487	(100.00)	127,463	(100.00)	489	(100.00)	128,322	(100.00)

Note. REG, regular; DTR, transparency doubtful; OPQ, opaque; FCO, formally complex only; ALM, allomorphy; PSE, pseudo-prefixed; T, total.

in serial search models with and without prefix stripping. In other words, we estimate the total sum of the number of search steps required for the processing of all word tokens that have an initial string matching any of the prefixes studied here. For both models we assume that the input is checked for the presence of prefixes. In the model with prefix stripping, the serial search proceeds on the basis of the stem, in the model without prefix stripping the search proceeds on the basis of the prefix, the assumption being that all words beginning with some given prefix are listed in the central system together. In other words, in the stem-based search model of Taft and Forster (1975) the potential prefix is stripped, following which the search proceeds on the

TABLE 2
TYPE AND TOKEN COUNTS FOR PREFIXES AND PSEUDOPREFIXES FOR ENGLISH ON THE BASIS OF THE COBUILD CORPUS AS AVAILABLE UNDER CELEX

		Orthography				Phonology			
Prefix	Category	Types	(%)	Tokens	(%)	Types	(%)	Tokens	(%)
be-	TR	139	(34.)	3,068	(03.)	139	(49.)	3,068	(03.)
	OP	14	(03.)	60	(00.)	14	(05.)	60	(00.)
	PS	261	(63.)	90,693	(97.)	132	(46.)	88,095	(97.)
	T	414	(100.)	93,821	(100.)	285	(100.)	91,223	(100.)
de-	TR	346	(27.)	1,474	(02.)	327	(29.)	1,168	(02.)
	OP	4	(00.)	7	(00.)	4	(00.)	7	(00.)
	PS	920	(72.)	60,088	(98.)	793	(71.)	60,315	(98.)
	T	1270	(100.)	61,569	(100.)	1124	(100.)	61,490	(100.)
en-	TR	239	(48.)	4,038	(20.)	239	(46.)	4,038	(11.)
	OP	10	(02.)	72	(00.)	10	(02.)	72	(00.)
	PS	254	(50.)	16,184	(80.)	274	(52.)	34,010	(89.)
	T	503	(100.)	20,294	(100.)	523	(100.)	38,120	(100.)
re-	TR	709	(35.)	23,529	(16.)	709	(34.)	23,529	(16.)
	OP	44	(02.)	1,237	(01.)	44	(02.)	1237	(01.)
	PS	1283	(63.)	119,140	(83.)	1331	(64.)	123,306	(83.)
	T	2036	(100.)	143,906	(100.)	2084	(100.)	148,072	(100.)
un-	TR	472	(71.)	21,506	(76.)	472	(71.)	21,506	(56.)
	OP	8	(01.)	76	(00.)	8	(01.)	76	(00.)
	PS	182	(27.)	6,857	(24.)	185	(28.)	16,919	(44.)
	T	662	(100.)	28,439	(100.)	665	(100.)	38,501	(100.)
in-	TR	628	(36.)	27,998	(23.)	628	(35.)	27,998	(23.)
	OP	12	(01.)	5,054	(04.)	12	(01.)	5,054	(04.)
	PS	1120	(64.)	88,009	(73.)	1130	(64.)	89,168	(73.)
	T	1760	(100.)	121,061	(100.)	1770	(100.)	122,220	(100.)
mis-	TR	270	(78.)	1,517	(16.)	270	(76.)	1,517	(15.)
	OP	15	(04.)	1,288	(13.)	15	(04.)	1,288	(12.)
	PS	60	(17.)	6,895	(71.)	69	(19.)	7,503	(73.)
	T	345	(100.)	9700	(100.)	354	(100.)	10,308	(100.)

Note. TR, transparent; OP, opaque; PS, pseudo-prefix; T, column total.

basis of the remaining stem (c.q. BOSS) in the access lexicon. In our base-line model no stripping takes place. As soon as a potential prefix is recognized, a serial search is initiated within the central lexicon in the bin where all (full) words beginning with that (true) prefix are listed in order of their frequencies. If no potential prefix is detected, a search is initiated in the access bin using some formal representation of the input, e.g., the first constituent, or the BOSS, as search key. The base-line model is designed to provide a serial search algorithm that is as similar as possible to the original model proposed by Taft and Forster (1975). The only difference is that prefixed words are now handled in exactly the same way as other morphologically complex words where access proceeds on the basis of the first constituent.

To analyze the two competing search algorithms we calculate estimates of the required number of search steps needed to successfully locate a word in the lexicon, where a search step is defined as the operation of comparing the search key of that word with a (single) entry in some bin. Since the length and composition of the ac-

cess and central bins varies with the search algorithms, a number of different variables have to be taken into account. For each of the two algorithms, the values of these variables are calculated using our lexical databases.

The first variable of interest is the length a of the phonological and orthographical access bins. This length is especially relevant in the case of pseudo-stems, where only after a search steps through the access bin it is discovered that the search key does not match any input entry. A second variable is the mean length of a bin in the central system. Here we have to make a distinction between the mean length of the central bins in the prefix stripping model c and the mean length of such bins in the base-line model, where we have the (generally large) mean length of the prefix bins p and the much smaller mean length of the stem-based or BOSS-based bins d.

Next we need to know the (expected) number of search steps required to locate a given item in a frequency ordered list. The expectation of the number of search steps required for a frequency-ordered list L, e_L, can be calculated as shown in the appendix. We have calculated the expected numbers of search steps for English and Dutch access bins both in terms of stems and using BOSS entries. No attempt has been made to calculate the mean expected number of search steps for the multitude of central bins. Rather than that, we computed the relative advantage of using a frequency-based search with respect to the length of the access bin, i.e., the maximal number of steps possible. In the case of the access bins, this relative advantage f equals e_L/a, the ratio of the expected number of search steps to the length of the bin. We then used this ratio to estimate the expected number of search steps for the central bins.[7] Hence the estimated number of search steps in a central bin of, e.g., length c is given by $f \cdot c$.

Finally, the numbers of words with the relevant morphological constituencies are relevant here. Obviously, we need to know the number of prefixed word tokens x and the number of pseudo-prefixed word tokens y. In addition, we need to distinguish between two types of pseudo-prefixed words, namely, type I where there is no existing stem (*decorum*) and type II where a semantically unrelated stem leads to erroneous searches in the central system (*begin*). We have calculated the proportion g of tokens of type I to the total number of pseudo-prefixed tokens y for both stem-based and BOSS-based representations.

Table 3 summarizes the values of these variables for English and Dutch orthographical and phonological wordforms using both stem-based and BOSS-based representations. The values for x and y can be directly obtained from Tables 1 and 2. The lengths of the stem-based access bins were obtained by extracting all stems from the CELEX database of morphologically parsed word types. The lengths of the BOSS-based access bins were derived by applying the definition of BOSS to all orthographical wordforms (including inflectional variants) and collapsing over identical strings. The mean lengths of the central bins c, d, and p were obtained by calculating for each stem and prefix the number of different word types that would appear in their respective bins and averaging over the numbers obtained. The calculation of f is explained in the appendix.

We are now in the position to estimate the number of search steps required by the different algorithms. In the case of the prefix stripping algorithm, we first consider the estimated number of search steps required for the access to truly prefixed words. The x tokens with a prefix require on average fa search steps in the access bin, hence a total number of $x \cdot fa$ search steps is involved.

[7] The exact mean expected number of search steps in the central system can be obtained only by actually building a lexicon organized along the lines of the Taft and Forster (1975) proposal.

TABLE 3
THE VALUES OF THE VARIABLES OF THE PREFIX STRIPPING MODEL AND THE BASE-LINE CONTROL MODEL FOR DUTCH AND ENGLISH ORTHOGRAPHICAL AND PHONOLOGICAL WORD FORMS

		Dutch		English	
		Stems	BOSS	Stems	BOSS
	a	5928	4436	11500	6679
	c	5.08	24	1.14	2.09
	d	4.77	23	1.14	2.55
	f	.06	.07	.07	.08
	e	325	320	802	547
	g	.67	.56	.67	.37
	p	608	608	413	413
Orth	*x*	1,289,668	1,289,668	90,924	90,924
	y	563,243	563,243	387,866	387,866
Phon	*x*	1,296,404		90,618	
	y	538,748		419,316	

Similarly, the estimated number of search steps in the relevant central bin is given by $x \cdot fc$.

Next consider the processing of pseudo-prefixed words of type I (*decorum*). There are gy such tokens. In the absence of an access code in the access bin, such tokens will require by definition a search steps in order to discover that the prefix should not have been stripped, hence $gy \cdot a$ search steps. In this case the access bin has to be re-accessed, this time using the full string as access code. Locating the matching entries for these gy tokens in the access bin will require $gy \cdot fa$ steps. In the central bin, $gy \cdot fc$ search steps are needed.

In the case of pseudo-prefixed words of type II (*begin*) the search key matches with an entry in the access bin. The $(1 - g)y$ tokens of this kind each require fa steps: a total of $(1 - g)y \cdot fa$ steps. For each of these $(1 - g)y$ tokens, an unsuccessful search in the relevant central bins is conducted, leading to $(1 - g)y \cdot c$ steps. Following this, the access bin is re-accessed with the full form: $(1 - g)y \cdot fa$ search steps. Finally the correct central bin can be searched, leading to a cost of $(1 - g)y \cdot fc$ steps. Summing up, the grand total of search steps t_{ps} required to complete lexical access for all prefixed and pseudo-prefixed words equals

$$t_{ps} = xf(a + c) + gya(f + 1) + (1 - g)y(2fa + c) + yfc. \quad [1]$$

The calculation of the required number of search steps for the base-line model is less cumbersome. Truly prefixed words require a single search in the relevant prefix bin in the central system. Hence the x prefixed tokens will take $x \cdot fp$ steps. The distinction between pseudo-prefixed words of type I and those of type II is not relevant here. In both cases a single exhaustive search through the relevant prefix bin in the central system will lead to the discovery that a full-form based search in the access bin is required. Since there are y pseudo-prefixed words the first scan will take $y \cdot p$ steps. The subsequent search in the access bin will take $y \cdot fa$ steps. Finally, $y \cdot fd$ steps are needed for locating these words in the central system. Summing up, the total number of required search steps t_{bl} equals

$$t_{bl} = xfp + y(p + f(a + d)). \quad [2]$$

Table 4 summarizes the results obtained when the data of Table 3 are plugged into these equations by listing the average number of search steps required for locating a

TABLE 4
THE ESTIMATED MEAN NUMBER OF SEARCH STEPS *s* REQUIRED BY THE PREFIX STRIPPING ALGORITHM AND THE ESTIMATED NUMBER OF SEARCH STEPS *b* FOR THE BASE-LINE ALGORITHM

		English stem	BOSS	Dutch stem	BOSS
ORTH	*s*	7262	2810	1599	1112
	b	992	774	318	309
	s/b	7.3	3.6	5.0	3.6
PHON	*s*	7360		1557	
	b	1007		309	
	s/b	7.3		5.0	

word token with a potential prefix string in the central system. What we find is that the prefix-stripping algorithm is clearly less efficient than the base-line model where words with the same prefix are stored together. A substantial increase in efficiency is obtained by the introduction of the BOSS, but even then the base-line model is roughly four times as efficient as the prefix stripping model. Observe that the prefix stripping model performs worse for English than for Dutch, but that modality does not make much of a difference for stem-based input systems. While the original motivation of the prefix stripping model was to *increase* the efficiency of the serial search by *not* storing prefixed words together, it now appears that the remedy is (much) worse than the disease.

It might be argued that the observed inefficiency of the prefix-stripping algorithm is an artifact of our definition of pseudo-prefixation. Would the algorithm fare better if the original description of pseudo-prefixation in Taft and Forster (1975, 1976) had been adopted? To answer this question, we recalculated the number of prefixed and pseudo-prefixed words for English orthographical words, using the following criteria for prefixation with bound stems (such as *-gress* in *regress*):

1. The *Random House Dictionary of English* or the *Webster's Third New International Dictionary* should acknowledge the presence of a prefix in its etymology;

2. In addition, any bound stem thus identified should occur in at least two different derived words. The second criterion ensures that words such as *repertoire,* pseudo-prefixed according to Taft and Forster (1975; p. 646), where the prefix *re-* is present etymologically (from Latin *re* + *parire*), are not considered to be prefixed: there are no other words with this putative bound stem.

As a result, not only fairly transparent forms like *-gress* and *-cline* are counted as bound stems, but also semantically opaque stems such as in *demise* and *promise* and *rebate* and *debate*. We have also counted etymologically unrelated words as prefixed when their stems are phonologically identical in present-day English: *remember* (from Latin *re* + *memor*) and *dismember* (from Old French *des* + *membrer*), *recant* (from Latin *re* + *cantare*) and *decant* (from Latin *de* + *cantus*), *relieve* (from Latin *re* + *levare*) and *believe* (from Middle English *bi* + *leven*). The net effect is to reclassify a substantial number of words originally classified as pseudo-prefixed.

The resulting type and token counts, which follow the original Taft and Forster proposals as closely as possible without any additional assumptions from our side, are listed in Table 5.

When we compare Table 5 with Table 2, we find that the ratios of prefixed to pseudo-prefixed word tokens have shifted considerably in favor of the prefixed words. Nevertheless, the overall error rate for the prefixes considered here remains high at 56% (compared to 81% in the original analysis). To evaluate the processing efficiency of the prefix stripping model with this definition of prefixation we recalculated the parameters of Table 3 for written words, obtaining a mean number of search steps equaling 7035 for the prefix stripping model and 886 for the base-line model. Thus it turns out that prefix stripping requires about 8 times (7.94) as many search steps. Even the application of the length criterion discussed above does not improve the effi-

TABLE 5
TYPE AND TOKEN COUNTS FOR PREFIXES AND PSEUDO-PREFIXES, USING A STRICTLY ETYMOLOGICAL DEFINITION OF PSEUDO-PREFIXATION, FOR ENGLISH ON THE BASIS OF THE COBUILD CORPUS AS AVAILABLE UNDER CELEX

		Orthography				Phonology			
Prefix	Category	Types	(%)	Tokens	(%)	Types	(%)	Tokens	(%)
be-	PR	186	(31.)	62,578	(33.)	187	(39.)	62,578	(52.)
	PS	422	(69.)	129,562	(67.)	294	(61.)	58,126	(48.)
	T	608	(100.)	192,140	(100.)	481	(100.)	120,704	(100.)
de-	PR	987	(65.)	52,143	(54.)	858	(35.)	43,880	(30.)
	PS	541	(35.)	44,375	(46.)	1569	(65.)	103,508	(70.)
	T	1528	(100.)	96,518	(100.)	2427	(100.)	147,388	(100.)
en-	PR	363	(65.)	8,225	(18.)	361	(67.)	8,191	(10.)
	PS	193	(35.)	37,097	(82.)	179	(33.)	75,958	(90.)
	T	556	(100.)	45,322	(100.)	540	(100.)	84,149	(100.)
re-	PR	1415	(65.)	69,900	(42.)	1455	(58.)	70,815	(39.)
	PS	749	(35.)	98,388	(58.)	1037	(42.)	112,226	(61.)
	T	2164	(100.)	168,288	(100.)	2492	(100.)	183,041	(100.)
un-	PR	498	(69.)	32,425	(48.)	492	(74.)	21,911	(56.)
	PS	228	(31.)	35,578	(52.)	175	(26.)	16,884	(44.)
	T	726	(100.)	68,003	(100.)	667	(100.)	38,795	(100.)
in-	PR	1188	(66.)	79,732	(65.)	1188	(66.)	79,732	(65.)
	PS	614	(34.)	43,403	(35.)	600	(34.)	43,089	(35.)
	T	1802	(100.)	123,135	(100.)	1788	(100.)	122,821	(100.)
mis-	PR	290	(84.)	2,821	(29.)	290	(82.)	2,821	(27.)
	PS	55	(16.)	6,879	(71.)	64	(18.)	7,487	(73.)
	T	345	(100.)	9,700	(100.)	354	(100.)	10,308	(100.)

Note. PR, prefixed; PS, pseudo-prefixed; T, column total.

ciency of the model. For instance, if we exclude a priori prefix stripping resulting in putative stems with one or two letters, which reduces the number of pseudo-prefixed tokens by roughly 100,000, the prefix stripping model is again found to require roughly 8 times as many search steps (7.8). Apparently, any reduction in the number of pseudo-prefixed words affects the prefix-stripping model and its base-line counterpart in roughly the same way.[8] For Dutch, similar calculations can be carried out by counting only the words explicitly tabulated as PSE in Table 1 as pseudo-prefixed, excluding the categories FCO and ALM that were considered as pseudo-prefixed in the original analysis. The percentages of pseudo-prefixed words now are 16.9% for written words and 15.3% for spoken words. For written words, the mean number of search steps is estimated at 944

[8] In addition, recall that any reduction in the percentage of pseudo-prefixed orthographical tokens brought about by ruling out shorter pseudo-prefixed words is offset by an increase in the number of short truly prefixed word tokens that are now *not* accessed by their stem. For instance, for English *de-*, exempting words with a length of 1–6 letters from prefix stripping will reduce the percentage of pseudo-prefixed tokens from 46 to 17%, at the cost of not parsing 8% of the tokens with a real prefix. It should be noted that, without any further ad hoc assumptions, these short prefixed words will induce additional processing costs similar to those caused by type 1 pseudo-prefixed words. These extra costs have not been taken into account in the efficiency calculations of Table 4, which, from this point of view, presents the upper boundary of search efficiency.

for the prefix stripping model and 214 for the base-line model. Again we find that a reduction in the number of pseudo-tokens affects the overall processing efficiency only marginally: the original factor of 5.0 in Table 4 reduces to 4.4. Similar results can be obtained for English and Dutch spoken words.

We are forced to conclude that Knuth's (1973) original insight which appears to have motivated the prefix stripping hypothesis cannot be carried over to psychological models of lexical processing. The distributional properties of the languages studied here show that the costs of this procedure outweigh the benefits.

General Discussion

We have presented a critical evaluation of the prefix stripping model by focusing on its computational aspects. We have analyzed the CELEX lexical database for the occurrence of words with prefixes and pseudo-prefixes in combination with their frequencies. It turned out there is a non-negligible number of tokens with pseudo-prefixes in Dutch and a substantial number of such word tokens in English.

Any token with a pseudo-prefix is associated with some cost in a prefix stripping model because it will lead to a processing error. We have estimated the total number of processing steps that would be needed to analyze all the tokens of words with a potential prefix. It turned out that prefix stripping requires 3 and up to 8 times more search steps than an alternative model where words with the same prefixes are stored together. The huge error rate and the higher processing costs of prefix stripping show that, at least for English and Dutch, this theory is misguided.

It must be noted, that there are no *inherent* problems with the proposal of prefix stripping as such. It is only the vocabularies and the frequencies with which certain words occur that are causing the observed problems for prefix stripping in a serial search scheme. Given another lexical orthographic/phonological space and other frequency data prefix stripping could in principle be a viable option. For English and Dutch, the introduction of prefix stripping fails to accomplish its intended purpose, the increase of processing efficiency.

Of course, the efficiency of the original search model can be significantly increased by speeding up the first phase of the access process, locating a word in an access bin. Such an increase can already be observed in our data when lemma-based access is compared with BOSS-based access. By using BOSSes rather than lemmas as access keys the length of the (orthographical) access bin is reduced considerably, thereby speeding up serial search. Another way of reducing the number of search steps necessary to locate a word in an access bin is to distribute the entries over a series of smaller access bins, as suggested by Forster (1989), using formal similarities between words as the criterion for bin assignment. Undoubtedly, the introduction of a large enough number of such smaller access bins will lead to a serial search model for which prefix stripping can be incorporated without loss of processing efficiency—although one risks to lose the cumulative frequency effect when too many bins are used (Forster, 1989:83). Paradoxically however, the introduction of such additional hash-coding machinery obviates the necessity of prefix stripping itself. Consider the original efficiency argument, which proceeded on the assumption that listing very large numbers of prefixed words under the same description should be avoided (Taft & Forster 1975, p. 645). Given adequate and efficient methods for dealing with serial searches in large bins, the necessity of avoiding such very large sets of prefixed words evaporates. Evidently, prefix stripping cannot be motivated on the basis of efficiency arguments.

It is instructive to briefly compare the prefix stripping theory with another theory that stipulates obligatory stripping of word-initial materials, namely, the metrical seg-

mentation strategy proposed in Cutler and Norris (1988), Cutler and Carter (1987), and Cutler and McQueen (in press). These authors argue that in the perception of continuous speech word boundaries are initially postulated to be located at the beginnings of metrically strong syllables. This implies that words beginning with metrically weak, that is, fully unstressed, syllables require some additional procedure for evaluating the role of the stripped weak syllable. For instance, the second syllable of *horizon* would be an access code along with e.g., *rise* and *riser* in the same cohort, to be matched against the input string. The stripped initial syllable should remain available for this matching process, in which the full input representation has to be compared with the complete (phonetic) representation of *horizon* that is listed under the access code *rizon*. Similarly, a prefixed word with an unstressed initial prefix such as *reforest* will be accessed via its stem, *forest*.

As in the prefix stripping theory, a question arises as to how words are processed for which the stripping of the initial segment is not supported by the morphology. Recall that Taft and Forster argued that pseudo-prefixed words require more processing steps than words with real prefixes. In other words, they claim that pseudo-prefixation induces extra processing costs with respect to true prefixation. In the metrical segmentation theory, there is no extra cost in terms of the number of processing steps associated with words such as *horizon*, as it is assumed that *horizon* is listed under *rizon*. Nevertheless, there may be some costs associated with having such linguistically rather ad hoc access codes. On the other hand, prefixed words with a weak initial syllable will probably be processed more quickly than words like *horizon* where no prefix is present: in the latter case, neither the stripped initial syllable nor the remaining word find any support as linguistically relevant units elsewhere in the language. In the former case, however, it may well be that skipping a metrically weak prefix will speed up recognition compared to words containing a stressed initial prefix: other things being equal, the cohort sizes of stems are substantially smaller than those of full, prefix initial words. Similarly, the uniqueness points of the stem-based access codes will be reached earlier than those of the access codes including the prefix. This suggests that an evaluation of the metrical segmentation strategy on the word level should proceed in terms of a comparison of the number of words with a weak initial syllable and an initial prefix with the number of weak initial words that do not contain a prefix.

Inspection of the CELEX lexical databases revealed that 91.76% of all English word tokens begin with a strong syllable, a percentage of the same order of magnitude as reported in Cutler and Carter (1987). Prefixed words beginning with a weak syllable account for 0.9% of all tokens in the database, and the words without a prefix but with an unstressed initial syllable for 7.35%. For Dutch, these percentages are 87.53, 7.49, and 4.97%, respectively. Thus we find that for the words with a weak initial syllable the ratio of the number of word tokens with a prefix to the number of tokens without a prefix is roughly 1:8 for English and 3:2 for Dutch. This suggests that for Dutch, but not for English, possible costs of the metrical segmentation strategy for words like *horizon* may already be counterbalanced on the word level by the advantages for words like *reforest* associated with stem-based access.[9] If this line of reasoning is correct, the distributional

[9] On the other hand, this comparison is complicated by the observation that in Dutch the number of words where after the stripping of a weak initial syllable an existing word remains (compare English *vermillion*) is substantially higher than in English. Cutler & McQueen (in press) calculated that 0.7% of all word tokens in their corpus are of this type. Our calculations for Dutch, however, point to 3.1% of all tokens. If one assumes, with Cutler & McQueen, that such words lead to garden pathing, this will occur more often in Dutch than in English.

properties of the English lexicon appear to be non-optimal for a segmentation strategy that capitalizes on metrically strong syllables. This does *not* imply, of course, that the lexical segmentation strategy is ruled out for English on statistical grounds. Briscoe (1989) presents some computational analyses suggesting that the metrical segmentation strategy performs better than strategies based on phoneme, syllable or word properties. This suggests that possible disadvantages of metrical segmentation at the word level may be outweighed by its merits on the sentence level. Moreover, there is accumulating experimental evidence that for English the metrical segmentation theory may well be correct (Cutler & Norris, 1988; Cutler & Butterfield, 1992; McQueen, Norris, & Cutler, 1994).

In sum, our lexical statistics suggest that if some form of early stripping of initial segments does take place, initial syllables are more likely candidates than initial (pseudo) prefixes.

Thus far we have been concerned with uncovering the quantitative distributional properties of prefixed and pseudo-prefixed words in English and Dutch. We have seen that the number of pseudo-prefixed words is a function of both prefix and language. This raises the question whether the presence of pseudo-prefixed words is in any way relevant to lexical processing. The answer appears to be yes. Laudanna and Burani (in press) studied a number of Italian prefixes varying both the number of pseudo-prefixed types and the number of such tokens in a series of visual lexical decision experiments. What they found is that response latencies increased for pseudowords containing a real prefix as a function of the number of pseudo-prefixed words. In nonwords, prefixes with little pseudo-prefixation required longer response latencies to reject than prefixes with substantial pseudo-prefixation. As the authors argue, this suggests that the presence of pseudo-prefixed words affects the availability of the prefix as a functional unit in word recognition. It still remains to be shown that similar results can be obtained for other languages. Even so, the present data indicate that one should avoid generalizing over prefixes as such. The distributional properties of prefixes may vary substantially from prefix to prefix. This will influence the way in which a given prefix functions in the mental lexicon. Some prefixed words may require on-line decomposition, for others this will not be the case. Models of morphological processing have tended to view decomposition as an all or nothing issue, however. Full listing models (Butterworth, 1983) have been formulated as well as strict parsing schemes (Pinker, 1991). Others have argued that (regular) inflectional morphology, but not (regular) derivational morphology, is decomposed (Friederici, Schriefers, & Graetz, 1989). The prefix stripping model evidences the same line of thinking: any initial string resembling a prefix is stripped, independent of the distributional properties of the prefix involved. Both our lexical statistics and Laudanna and Burani's (in press) experimental evidence suggest that this is too rigid a way of theorizing. Similarly, in an approach along the lines suggested by Cutler & Norris (1988) the critical properties of prefixes may vary considerably, both with respect to the metrical strength of the prefix (some prefixes are always unstressed, others always stressed, and some may be both stressed or unstressed, depending on the metrical structure in which they appear) and with respect to its distributional properties. Again it is hazardous to generalize over prefixes as an undifferentiated category. A model in which the properties of individual affixes crucially determine the behavior and the development of the lexical access system is outlined in Schreuder & Baayen (in press).

Finally, the main methodological point of our exercise is the following. In the same way as researchers confront the predictions of their models with subjects' experimental performance, one should also confront

one's model with the words as they appear in the language. Metaphorically speaking, the model has to be confronted with the (cruel) environment in which it has to perform. From this point of view, models can be tested empirically in at least two ways, namely by testing empirical predictions of the model with respect to behavioral data, and by actually letting the model do what it was designed for, using a realistic sample of language input. This double testing may lead to a dilemma—what happens when one test is favorable to the model and the other is not? It could be argued that we have such a case here, because there are a number of word recognition studies that claim to have found evidence for prefix stripping (e.g., Taft, 1981). The evidence, however, is controversial (see e.g., Henderson, 1985; Taft, 1992). If indeed some form of early parsing of prefixed words may occur under certain circumstances for a specific prefix, then the results of the present study show that one should look for other models of lexical processing (e.g., Caramazza, Laudanna, & Romani, 1988; Frauenfelder & Schreuder, 1992; Baayen, 1993, Schreuder & Baayen, in press, Baayen & Schreuder, in press), models without across the board parsing for all prefixes.[10]

Summing up, re-revisiting prefix stripping has shown that it is highly improbable that prefix-stripping as envisioned by Taft and Forster (1975) is involved in lexical processing. Furthermore, we have shown that a vocabulary study of words in relation to their frequencies of use may yield valuable empirical information essential to constraining models of word recognition.

Appendix

Let X be the required number of search steps for locating some word token in a frequency ordered list L. The expected number of search steps E[X] in such a list is obtained as follows. Let k be the length of L. Given that $fr + 1 \geq fr$, we have that E[X] is given by

$$E[X] = \frac{1}{N} \sum_{r=1}^{k} rf_r.$$

To see this, note that the most frequent word, having rank 1, will require a single search step. The second most frequent word will be located in 2 steps, etc.:

rank	frequency	# comparisons	
1	f_1	$1 \cdot f_1$	$1f_1$ tokens are found in 1 step
2	f_2	$2 \cdot f_2$	$2f_2$ tokens are found in 2 steps
3	f_3	$3 \cdot f_3$	$3f_3$ tokens are found in 3 steps
k	f_k	$k \cdot f_k$	kf_k tokens are found in k steps

Observe that the total number of search steps required for processing all N tokens of varying frequency is given by

$$\sum_{r=1}^{k} r \cdot f_r.$$

The probability that a word token with frequency f_r is encountered and has to be located in L equals

$$\Pr(X = r) = \frac{f_r}{N},$$

hence

$$E[X] = \sum_{r=1}^{k} r \frac{f_r}{N} = \frac{1}{N} \sum_{r=1}^{k} rf_r.$$

The ratio f of the expected number of search steps to the maximally possible number of search steps k is

$$f = \frac{\mathrm{E}[X]}{k}.$$

[10] Obligatory parsing is more likely to be involved in the processing of suffixed words, where one finds extremely low numbers of pseudo-suffixed words and where higher degrees of semantic transparency are more likely to be observed (see Baayen, 1993).

References

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Dordrecht: Kluwer.

Baayen, R. H. (1993). On Frequency, transparency and productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1992* (pp. 227–254). Dordrecht: Kluwer.

Baayen, R. H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics* 29, 801–843.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database*. (CD-ROM).

Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

BAAYEN, R. H., & SCHREUDER, R. (in press). Modeling the processing of morphologically complex words. In A. Dijkstra & K. de Smedt (Eds.), *Computational psycholinguistics: Symbolic and subsymbolic models of language processing*. Hemel Hempstead: Harvester Wheatsheaf: Simon and Schuster International Group.

BERGMAN, M. W., HUDSON, P. T. W., & ELING, P. A. T. (1988). How simple complex words can be: Morphological processing and word representation. *Quarterly Journal of Experimental Psychology*, **40A**, 41–72.

BRADLEY, D. C., & FORSTER, K. I. (1987). A reader's view of listening. *Cognition* **25**, 103–134.

BRISCOE, E. J. (1989). Lexical access in connected speech recognition. *Proceedings of the 27th Congress, Association for Computational Linguistics* (pp. 84–90). Vancouver, British Columbia, Canada.

BURNAGE, G. (1990). *CELEX. A guide for users*. Nijmegen: CELEX.

BURANI, C., & CARAMAZZA, A. (1987). Representation and processing of derived words. *Language and Cognitive Processes*, **2**, 217–227.

BUTTERWORTH, B. (1983). Lexical Representation. In B. Butterworth (Ed.), *Language production (vol. II): Development, writing and other language processes* (pp. 257–294). London: Academic Press.

CARAMAZZA, A., LAUDANNA, A., & ROMANI, C. (1988). Lexical access and inflectional morphology. *Cognition*, **28**, 297–332.

CHITASHVILI, R. J., & BAAYEN, R. H. (1993). Word frequency distributions. In G. Altman & L. Hřebíček (Eds.), *Quantitative Text Analysis*, 54–135. Trier: Wissenschaftliche Verlag.

COLÉ, P., BEAUVILLAIN, C., & SEGUI, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language*, **28**, 1–13.

Collins Dictionary of the English Language. (1986). London/Glasgow: Collins.

CORBIN, D. (1985). Comment intégrer l'exception dans un modèle lexical [How to integrate the exception in a lexical model]. *Langue Française* **66**, 54–76.

CUTLER, A., & BUTTERFIELD, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, **31**, 218–236.

CUTLER, A., & CARTER, D. M. (1987). Strong initial syllables in English. *Computer Speech and Language*, **2**, 133–142.

CUTLER, A., & MCQUEEN, J. M. (in press). The recognition of lexical units in speech. In B. de Gelder & J. Morais (Eds.), *From Spoken to Written Language*. Cambridge, MA: The MIT Press.

CUTLER, A., & NORRIS, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113–121.

FORSTER, K. I. (1989). Basic issues in lexical processing. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process* (pp. 75–108). Cambridge, MA: the MIT Press.

FRAUENFELDER, U. H., BAAYEN, R. H., HELLIWIG, F. M., & SCHREUDER, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, **32**, 781–804.

FRAUENFELDER, U. H., & SCHREUDER, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 165–183). Dordrecht: Kluwer.

FRIEDERICI, A. D., SCHRIEFERS, H. J., & GRAETZ, P. A. (1989). Abruf und Repräsentation morphologisch komplexer Wörter verschiedener Wortklassen [Access and representation of morphologically complex words of various word classes]. In H. Guenther (Ed.), *Experimentelle Studien zur deutschen Flexionsmorphologie*, Hamburg: Helmut Buske Verlag.

HENDERSON, L. (1985). Towards a psychology of morphemes. In A. W. Ellis (Ed.), *Progress in the Psychology of Language, Vol. 1. (pp. 15–72)*. London: Erlbaum.

KNUTH, D. E. (1973). *The Art of Computer Programming. Vol. 3: Sorting and Searching*. Reading, MA: Addison-Wesley.

LAUDANNA, A., & BURANI, C. (in press). Distributional properties of derivational affixes: implications for processing. In L. Feldman (Ed.), *Morphological Aspects of Language Processing*. Hillsdale, NJ: Erlbaum.

LIMA, S. D. (1987). Morphological analysis in sentence reading. *Journal of Memory and Language*, **26**, 84–99.

MANELIS, L., & THARP, D. A. (1977). The processing of affixed words. *Memory & Cognition*, **5**, 690–695.

MCQUEEN, J. M., NORRIS, D., & CUTLER, A. (in press). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

PINKER, S. (1991). Rules of language. *Science*, **253**, 530–535.

RENOUF, A. (1987). Corpus development. In J. M. Sinclair (Ed.), *Looking up: An account of the cobuild project in lexical computing* (pp. 1–40). London: Collins.

SCHREUDER, R., & BAAYEN, R. H. (in press). Modelling morphological processing. In L. Feldman

(Ed.), *Morphological aspects of language processing*. Hillsdale, NJ: Erlbaum.

SCHRIEFERS, H., ZWITSERLOOD, P., & ROELOFS, A. (1991). The identification of morphologically complex spoken words: Continuous processing or decomposition? *Journal of Memory and Language*, **30**, 26–47.

The Shorter Oxford English Dictionary. (1973). Oxford: The Clarendon Press.

TAFT, M. (1979a). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, 7, 263–272.

TAFT, M. (1979b). Lexical access via an orthographic code: The basic orthographic syllabic structure (BOSS). *Journal of Verbal Learning and Verbal Behavior*, **18**, 21–39.

TAFT, M. (1981). Prefix stripping revisited. *Journal of Verbal Learning and Verbal Behavior*, **20**, 289–297.

TAFT, M. (1985). The decoding of words in lexical access: A review of the morphographic approach. In D. Besner, T. G. Waller, & G. E. Mackinnon (Eds.), *Reading research: Advances in theory and practice* (pp. 83–123). London: Academic Press.

TAFT, M. (1988). A morphological decomposition model of lexical representation. *Linguistics*, **26**, 657–667.

TAFT, M. (1992). *Reading and the Mental Lexicon*. Hove: Erlbaum.

TAFT, M., & FORSTER, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, **14**, 638–647.

TAFT, M., & FORSTER, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, 15, 607–620.

TAFT, M., HAMBLY, G., & KINOSHITA, S. (1986). Visual and auditory recognition of prefixed words. *Quarterly Journal of Experimental Psychology*, **38A**, 351–386.

VRIES, J. W. DE (1975). *Lexicale morfologie van het werkwoord in modern Nederlands. (Lexical morphology of the verb in modern Dutch.)* Leiden: Universitaire Pers.

Webster's Third New International Dictionary of the English Language. (1981). Springfield, MA: Merriam–Webster, Inc.

(Received December 31, 1992)

(Revision received September 1, 1993)