LANGUAGE AND COGNITIVE PROCESSES, 1994, 9 (3) 447–469

Productivity in Language Production

R. Harald Baayen

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Lexical statistics and a production experiment are used to gauge the extent to which the linguistic notion of morphological productivity is relevant for psycholinguistic theories of speech production in languages such as Dutch and English. Lexical statistics of productivity show that despite the relatively poor morphology of Dutch, new words are created often enough for the marginalisation of word formation in theories of speech production to be theoretically unattractive. This conclusion is supported by the results of a production experiment in which subjects freely created hundreds of productive, but only a handful of unproductive, neologisms. A tentative solution is proposed as to why the opposite pattern has been observed in the speech of jargonaphasics.

INTRODUCTION

The present study addresses the question of whether the linguistic notion of morphological productivity, the ability of speakers to coin, without any apparent conscious effort, new morphologically complex words, is relevant for theories of lexical access in language production. Butterworth (1983) suggested that the procedures for dealing with neologisms are, at least in languages like English or Dutch, of an analogical nature, crucially depending on some not fully predictable lexical item serving as the model for the creation of a new form. The high proportions of neologisms formed on the basis of a non-productive word formation pattern in the speech of three Italian jargonaphasics (Semenza, Butterworth, Panzeri, & Ferreri, 1990) can be interpreted as evidence in support of Butterworth's position. Levelt (1989) similarly argues that for a language like English, the passive store of

Requests for reprints should be addressed to R.H. Baayen, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. E-mail: baayen@mpi.nl.

I am indebted to Dominiek Sandra, Rob Schreuder, Pim Levelt, Ardi Roelofs and Jos van Berkum for valuable discussion.

declarative knowledge is of primary importance, with procedural knowledge (rule-governed word formation) playing a negligible role.

In this paper, both lexical statistics and psychological experimentation are used to gauge the extent to which the linguistic notion of morphological productivity is relevant for psycholinguistic theories of speech production in Dutch. The results obtained suggest that even though Dutch morphology is indeed of marginal interest and importance compared with the rich derivational and inflectional morphological systems one may observe in, for example, Altaic or Eskimo-Aleut languages, novel morphologically complex forms occur too often for the marginalisation of procedural lexical knowledge to be attractive theoretically.

Our discussion is structured as follows. The next section outlines a number of ways in which we may quantify the productivity of word formation processes on the basis of text corpora. Focusing on a number of Dutch affixes, clear predictions of the likelihood of the realisation of neologisms can be made. These predictions are tested in the following section, using an experimental technique developed by Anshen and Aronoff (1988). The implications of the results obtained—subjects appear to be able to effortlessly coin hundreds of neologisms with productive affixes—are discussed in the final section.

LEXICAL STATISTICS AND PRODUCTIVITY

How often do speakers of languages such as English or Dutch encounter new words? Let us assume that a speaker with a normal speech rate will utter approximately 150 words per minute, as suggested by Maclay and Osgood (1959). Also assume that one is engaged in speaking (and listening) for some 12 hours a day. If so, one will use some 108,000 word tokens per day, or 39,420,000 word tokens per year. This is roughly the size of a corpus of modern Dutch compiled by the Dutch Institute for Lexicography (INL). Assuming that words encountered only once a year have no lexical representation of their own in the mental lexicon,[1] and hence have to be created on-line on the basis of one's procedural knowledge, we can estimate the daily number of morphologically complex words requiring productive word formation on the basis of the words occurring once only

[1]This assumption is not unrealistic in light of the fact that many monomorphemic words with a relative frequency of less than 10/42,000,000 are unknown to all but those who happen to be "experts" in the semantic domain to which such words belong.

(the so-called hapax legomena) in this corpus of 42 million word forms.[2] What we find is that roughly 33 hapaxes in a lemma-based analysis and 62 hapaxes in a wordform-based analysis are encountered every day. These estimates suggest that, even though the role of morphological procedural knowledge may be marginal token-wise—only 0.03% of the word tokens encountered daily require productive word formation—the absolute number of words requiring some form of productive word formation is non-negligible.

For English, similar results can be obtained. For instance, readers of *The Times* encountered some 30 formations in *-ness* in January 1993 that had not been used in any issue of *The Times* between September 1989 and December 1992. Roughly half of these hapaxes in the 90 million corpus of *The Times* are not registered in *Webster's Third International Dictionary* (1981), nor in Merriam's (1981) supplement to this dictionary. For this one suffix, already one or two new formations are used every other day (Renouf & Baayen, 1994), suggesting a substantially larger number of daily neologisms for all derivational and inflectional affixes jointly.

The present estimates of the rate at which new complex words are encountered in English and Dutch should be interpreted with caution, however, especially in light of the differences in vocabulary richness between spoken and written language. In spontaneous speech, the rate at which novel complex words are produced is much lower than for carefully composed written prose. The same holds for listening versus reading. To complicate matters even further, there is an asymmetry between the relatively poor language output of the individual and the rich output of the language community as a whole in which the same individual is immersed. In fact, individual speakers or writers may be reluctant to use particular words or word formation processes, even though they have no problems comprehending these words in the speech or texts of others. The precise way in which comprehension and production, as well as the language

[2]These counts are based on the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993; Burnage, 1990). Unfortunately, the CELEX lexical database has not registered the hapaxes occurring in the INL corpus. Hence the numbers of hapaxes presented here are estimated lower bounds using the property (Muller, 1979, pp. 458–459) of word-frequency distributions in the LNRE ZONE (Chitashvili & Baayen, 1993) that:

$$V_N(m) > \frac{V_N(m+1)}{V_N(m+2)} V_N(m+1),$$

where $V_N(m)$ denotes the number of types occurring m times in a sample of size N. Similarly, the number of morphologically complex hapaxes was estimated using the proportion of morphologically complex words among the dislegomena, again resulting in a lower bound.

community and individual preferences, interact in shaping one's morphological competence is unknown. The estimates presented above give an upper bound for the rate at which individual speakers hear and produce new morphologically complex forms.

With this caveat in mind, we now turn to consider the phenomenon that word formation rules may differ with respect to their degree of productivity. Following Schultink (1961), I will take the notion "morphological productivity" to denote the possibility for language users to coin, unintentionally, a number of formations that is, in principle, enumerably infinite. This possibility, however, is not exploited to the same extent for different affixes. This is already apparent from the widely varying numbers of words with given affixes one may observe in standard dictionaries. Dictionary-based counts, however, are not the most reliable way to gauge the productivity of word formation rules. More precise results can be obtained in two ways.

The first is to investigate word use in very large text corpora, such as the newspaper corpora that are becoming available on CD-ROM. These collections of daily issues, often comprising tens of millions of tokens, can be scanned for the use of neologisms or very low-frequency items. For the de-adjectival nominalising suffixes *-ness* and *-ity*, such a scan of *The Times* as it appeared in 1991–92 reveals that some 15 neologisms a month in *-ness* (where a neologism is defined as a word not listed in the Webster's dictionary) are counterbalanced by only 5 neologisms in *-ity*. Evidently, *-ness* is the more productive suffix of the two (Renouf & Baayen, 1994).

Another way of gauging the degree of productivity of a word formation rule is to study the frequency distribution of the complex formations with that suffix in some representative corpus, using a probabilistic operationalisation of the notion "degree of productivity" along the lines of Bolinger (1948), who defined this notion as the statistically determinable readiness with which an affix enters into new combinations. The statistic that is most suited to measuring degrees of productivity is the so-called growth rate of the vocabulary, defined as the (mathematical) expectation of the ratio of the types that occur once only in a corpus or text, the hapax legomena, to the total number of tokens counted for that corpus or text (Baayen, 1989; 1992; 1993b; Baayen & Lieber, 1991). Denoting the hapaxes by $V_N(1)$ and the total number of word tokens counted in the corpus or text under consideration by N, the growth rate P of the vocabulary is estimated by the simple expression:

$$P = \frac{V_N(1)}{N} \tag{1}$$

The statistic P can be interpreted as the probability of encountering a word

type that has not yet been observed in the corpus or text under consideration.

This result from probability theory can be applied in two ways to the study of morphological productivity. We may consider the growth rate of the vocabulary as a whole, taking all words that occur in the corpus into account. The degree of productivity of a morphological category is then gauged by focusing on the contribution of this category to the overall growth rate. This can be shown to lead to what I have called the "hapax-conditioned degree of productivity" (Baayen, 1993b):

$$P^* = \frac{V_N(1,c)}{h}, \tag{2}$$

where $V_N(1,c)$ denotes the number of hapaxes belonging to the morphological category c under consideration, and h the total number of hapaxes in the corpus, irrespective of their morphological constituency. The interpretation of P^* pertains to the situation that the corpus is enlarged with one additional word token. What the statistic P^* estimates is the conditional probability that this additional token will belong to morphological category c given that this word token represents a new word type.

Another way of making use of formula (1) is to consider the growth rate of the morphological category itself. We now select all words belonging to morphological category c. We count the number of such word tokens N_c and the number of unique types $V_N(1,c)$. Using equation (1) we obtain what I have called the "category-conditioned degree of productivity":

$$P = \frac{V_N(1,c)}{N_c}. \tag{3}$$

The statistic P estimates the probability of sampling a word token that is new given that this word belongs to morphological category c.

For assessing the overall or "global" productivity (Baayen & Lieber, 1991) of some word formation process, the information provided by these productivity statistics, which pertain to the rate at which "new" types (types that have not been sampled before—mostly neologisms in large enough corpora) are expected to be encountered, should be supplemented by information concerning the extent to which the morphological processes involved have already led to "existing" formations. If two affixes appear with similar growth rates in a corpus, but the first shows up with twice as many types as the second, the first affix evidently is the more productive one in some global sense. We therefore take into account both the degree of productivity P or P^* and the number of "existing" types V_N, the "vocabulary" realised by the morphological process when N tokens have been sampled.

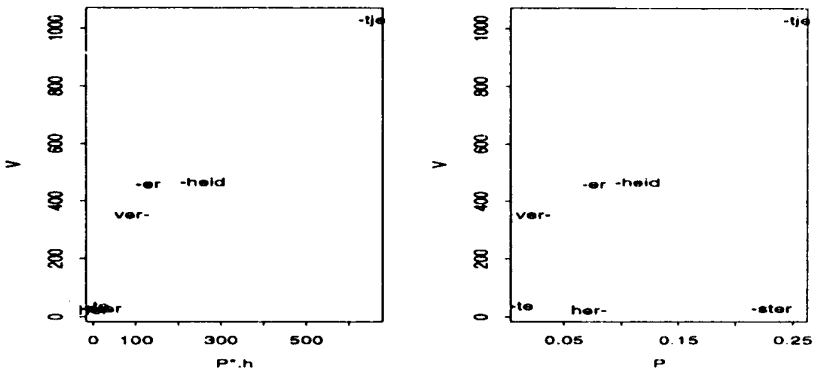

FIG. 1 Hapax-conditioned (left) and category-conditioned (right) degree of productivity plotted against the number of types V_N for selected Dutch affixes in the Eindhoven corpus.

Figure 1 summarises the global productivity of seven Dutch affixes: the diminutive suffix *-tje* (*laa-tje*, "small drawer"); the nominalising suffixes *-heid* and *-te*, which form abstract nouns from adjectives (*snel-heid*, "quickness, speed"; *warm-te*, "warm-th"); the nominalising suffixes *-er* and *-ster*, which form subject nouns and female subject nouns, respectively (*schrijver*, "writer"; *schrijf-ster*, "female writer"); and the verbalising prefixes *her-* (*her-evalueer*, "re-evaluate") and *ver-* (*ver-geef*, "forgive"). The degree of productivity is plotted on the horizontal axis; in the case of the hapax-conditioned degree of productivity (2), rescaled such that the horizontal axis shows the number of hapaxes $V_N(1,c)$. The realised vocabulary is plotted on the vertical axis. The corpus underlying these statistics is the Eindhoven corpus (Uit den Boogaart, 1975). Note that the extremely productive diminutive suffix *-tje* appears in the upper right-hand corner of both figures. It is characterised by a high number of realised types and a high growth rate. In contrast, the unproductive suffix *-te* appears in the lower left-hand corner. It is represented by only 39 types and shows up with a minimal growth rate, in contrast to its rival *-heid*, for which both V_N and $\boldsymbol{P}$ are fairly large. The two figures show a similar pattern for the affixes *-tje*, *-er*, *-heid* and *-te*. With respect to *her-* and *-ster*, however, they diverge. Their category-conditioned degree of productivity $\boldsymbol{P}$ (right-hand panel in Fig. 1) is much larger than their hapax-conditioned degree of productivity $\boldsymbol{P}^*$ (left-hand panel in Fig. 1). The question with which we shall be concerned is how this difference should be evaluated and interpreted.

To understand what is at issue here, recall that the hapax-conditioned degree of productivity $\boldsymbol{P}^*$ of some morphological category c can be

interpreted as the relative contribution of this category to the growth rate of the vocabulary. Like P^*, this relative contribution is a function of the sample size. In fact, it is known that the relative contribution of monomorphemic words to the growth rate of the vocabulary as a whole is a decreasing function of the sample size (Chitashvili & Baayen, 1993). Conversely, the relative contribution of productive affixes like English *-ly* or *-ness* to the growth rate of the vocabulary is an increasing function of the sample size. For large samples such as the Dutch INL corpus (42,000,000 wordforms), the relative contribution of morphology to the growth rate of the vocabulary at $N = 42{,}000{,}000$ has increased to 96% of the roughly 13,360 hapaxes in a lemma-based count.[3] In fact, as the sample size increases, P^* becomes an increasingly accurate estimate of the number of neologisms (defined with respect to some comprehensive dictionary) that may be expected to be found when the sample is enlarged. Similarly, given our assumption that words occurring once only in very large samples ($N \geqslant 40{,}000{,}000$) are not stored in the mental lexicon, we may use the hapax-conditioned degree of productivity P^* to gauge the extent to which different word formation rules are involved in the creation and perception of complex words.

The left-hand panel of Fig. 1 shows that suffixes such as *-heid* and *-tje* influence the growth rate of the vocabulary to a far greater extent than affixes such as *-ster* and *her-*. The former give rise to large numbers of neologisms and will often require lexical procedural knowledge in comprehension and production. The reverse holds for *-ster* and *her-*. Nevertheless, reference grammars of Dutch, such as that of Geerts, Haeseryn, de Rooij and Van den Toorn (1984), claim that *-ster* and *her-* are productive, in contrast to *-te*, a suffix with a similarly low P^* value. Apparently, the hapax-conditioned degree of productivity fails to tease apart the in some sense productive suffixes *her-* and *-ster* from unproductive *-te*. The method of counting neologisms in a large corpus shares the same deficiency: affixes such as *-ster* and *her-* will give rise to very few neologisms; nevertheless, they are known to be productive.

At this point, it is interesting to observe that the category-conditioned degree of productivity P assigns much higher degrees of productivity to *-ster* and *her-*. In fact, the P values are so high that they are rejected as unreliable by van Marle (1992). Does the P-based prediction, that in some sense *-ster* is likely to give rise to more neologisms than its unmarked counterpart *-er*, have any validity? Similarly, is stylistically marked *her-* in some sense more productive than stylistically unmarked *ver-*?

[3]On average, this amounts to one morphologically complex hapax in every 3281 word tokens in this corpus, or 33 every day in a lemma-based analysis. When inflectional variants of a lemma are counted as different types, this ratio increases to 1:1750, or 62 every day.

At first sight, these predictions seem counterintuitive, given that far more word types have been realised for *-er* and *ver-* than for *-ster* and *her-*. Moreover, the probability that one will hear or produce a new word in *-ster* or *her-* is quite low, as evidenced by P^*. In what sense, then, should we interpret the category-conditioned degree of productivity? Recall that P estimates the likelihood that a new type is used given that the morphological constituency of that type is already known. Since the reader/listener has no control over the speech input, it is hard to envisage how P might be relevant to speech perception. There is a stage in the production process, however, where P has a natural interpretation. Consider Levelt's (1989) production model, according to which the conceptual system generates a pre-verbal message that is the input to a formulating system that grammatically encodes the pre-verbal message. The process of grammatical encoding crucially depends on lexical knowledge. There are two kinds of lexical knowledge which may be exploited here: passive declarative knowledge and active, rule-governed procedural knowledge. Let us suppose that the conceptualiser has generated a pre-verbal message requiring the formulator to grammatically encode the lexical conceptual structure of some female personal agent noun. The statistic P can be interpreted as the probability that at this point in the production process, lexical procedural knowledge will have to be invoked to meet the requirements of the conceptualiser. Hence P is a more restricted productivity measure than the hapax-conditioned degree of productivity P^*. With respect to language production, the latter measure estimates the likelihood of the co-occurrence of the following events: (1) the conceptualisation process will require the grammatical encoding of a female agent noun, and (2) that the grammatical encoder will have to invoke procedural lexical knowledge in order to meet the requirements of the conceptualiser.

The first event is determined by a series of factors ranging from usefulness and stylistic appropriateness to conceptualisability and fashion. The second event is determined by the semantic transparency of the affix, the frequencies of the already existing items, the presence of rival affixes, the frequency with which the procedural knowledge specific to some affix has had to be invoked in the past, etc. In other words, while P^* evaluates the overall probability that a particular affix will be used or, alternatively, heard or read, P is specific to the process of grammatical encoding and may be expected to be especially sensitive to the extent to which the productivity of a category is muted by lack of semantic transparency, the presence of other ways of expression, affixed homonymy, etc. Also note that the divergence between the two productivity statistics P^* and P for the affixes *her-* and *-ster* mentioned above (see Fig. 1), suggests that a word formation process as such may be quite alive in the language (as indicated by a not too low value of P), even though it is called upon relatively seldom (as indicated by a low value of P^*).

Is it possible to create an experimental situation in which the conditions under which ***P*** is naturally interpretable are met? The possibility I will explore is that a production task developed by Anshen and Aronoff (1988) can be used here. Anshen and Aronoff asked subjects to write down as many words as they could think of in, for example, the English suffixes *-ness* and *-ity*, within 90 sec. Since this task forces a particular lexical conceptual structure upon the subjects, factors such as fashion, usefulness, or stylistic appropriateness—factors which normally co-determine the shape of the pre-verbal message—will generally have less force, or may even be absent. Hence we may expect the task to be especially sensitive to factors that determine grammatical encoding, such as the complexity of the lexical conceptual structures to be created, the frequencies of the existing complex words in the lexicon and the potential interference of alternative ways of expression.

What strategies can subjects exploit to meet the requirements of this task? Subjects may initiate a form-based scan, in the case of a suffix a scan similar to scanning the lexicon for words rhyming with that suffix.'Various errors due to affixal homonymy and pseudo-affixation may be expected to arise (Strategy 1). Subjects may also attempt a meaning-driven scan, as the target affix provides two kinds of semantic information that can be exploited. It specifies the word category and the semantic type of the complex word, and it delimits a semantic set of base words to which it can attach. One strategy subjects may pursue is to scan their lexicons for existing complex words using the categorial value of the affix as a search key (Strategy 2). Another strategy is to search for the appropriate base words, or, more precisely, the corresponding concepts. Once found, such a concept may drive lemma retrieval and subsequent phonological encoding. If the complex word happens not to exist, it can be created provided that the affix is productive (Strategy 3). Strategy 3 is less restricted than Strategy 2. For instance, when we ask subjects to coin nouns in *-heid*, they may use the second strategy and retrieve words from the relatively small set of abstract nouns, a set that does not function as such in the language system. But it is far easier to scan the mental lexicon for suitable adjectives, a major word category with many more members, and to use these adjectives to drive lemma retrieval. This suggests that subjects will prefer Strategy 3 to Strategy 2. We shall see that there is evidence that subjects use Strategies 1 and 3. Strategy 2 may have been used, but specific effects distinct from those of Strategy 1 could not be observed. Finally, note that if an affix is unproductive, subjects may tend to prefer form-based scanning (Strategy 1) to meaning-driven scanning (Strategy 3), in which case large numbers of form-based errors are to be expected.

Summing up, if we have analysed Anshen and Aronoff's production task correctly, and if the productivity measures discussed here have been interpreted correctly, we should find that:

1. Affixes with a high category-conditioned degree of productivity give rise to more neologisms than affixes with a low ***P*** value.
2. If subjects use form-based scanning (Strategy 1), erroneous responses involving pseudo-affixed forms and words with homonymic affixes will be produced, especially in the case of less productive affixes.
3. If subjects make use of all three strategies throughout the experiment, neologisms may be expected from the very first trial onwards due to Strategy 3.

A PRODUCTION EXPERIMENT

Method

Materials. To test these predictions experimentally, the following suffixes were selected for presentation:

1. The de-adjectival noun-forming rival suffixes *-te* and *-heid*, the former of which is unproductive or marginally productive at best (Baayen, 1989; 1992). If morphological productivity is relevant to the production task, more neologisms should be generated for *-heid* than for *-te*. Moreover, *-te* is expected to give rise to many errors.
2. The deverbal suffixes *-er* (unmarked) and *-ster* (marked), used to coin agent nouns. The inclusion of this pair allows us to check whether the category-conditioned degree of productivity correctly predicts that subjects should coin more new (marked) female agent nouns in *-ster* than unmarked agent nouns in *-er*.[4]
3. The prefixes *her-* and *ver-*. Even though the use of *her-* is restricted to more formal speech registers, it is semantically completely transparent. In contrast, the prefix *ver-* is realised in a substantial number of different formations, but its semantics are far more complex and less transparent (see Lieber & Baayen, 1994). The category-conditioned degree of productivity predicts that it is the more transparent of the two that should give rise to the larger number of neologisms.

Given the broad range of different factors that co-determine the productivity of a word formation process, we will in the main restrict ourselves to

[4]The suffix *-er* appears in a number of different forms: *-er*, *-der* and *-aar*. Following van Marle's (1985) suggestion that nouns in *X-der* and *X-er* jointly constitute a single morphological category distinct from that of nouns of the form *X-aar*, the counts to be presented below are restricted to personal nouns of the form *X-(d)er* and *X-(d)ster*, nominalisations in *X-aar* and *X-aar-ster* being left out of consideration.

comparing the experimental results for each of these three affix pairs separately, the affixes of these pairs being matched for the word category of the base, for the word category of the derived word, for affix type (prefix, suffix) and, in the case of the pairs *-heid/-te* and *-er/-ster*, to some extent for meaning. This will allow us to detect more precisely what factors are shaping the productivity of these affixes.

Procedure. Thirty-six subjects were paid to generate words for these six affixes as well as for the prefix *ge-*, which was used in a trial session to familiarise the subjects with the experiment. The written instructions contained examples of words with these affixes. The subjects were explicitly instructed that they were not required to produce only those words which they were sure would be listed in a dictionary. At the same time, they were asked to make sure that they only responded with words that did not violate the grammar of Dutch.

The affixes were presented in blocks. After a trial session with the prefix *ge-*, the subjects were asked to type into the computer words with the suffixes *-te* and *-heid*, the prefixes *ver-* and *her-*, and the suffixes *-er* and *-ster*. An incomplete counterbalanced design was used to eliminate statistically any possible effects of, for instance, experience gained early on in the experiment with the production of nouns in *-er* being carried over to the production of *-ster* words later on in the experiment.

For each of these affixes, the procedure was as follows. The target affix appeared towards the upper middle part of the computer screen, accompanied by a short sound signal. At the same time, a prompt was made available at the middle left-hand side of the screen. The subjects were asked to type in their words one at a time, completing each word with the RETURN key. The use of the RETURN key triggered the removal of the word from the screen. This removal was motivated by the wish to eliminate list effects as far as possible. After 5 min, the target affix at the top of the screen was replaced by the word "PAUZE" and further inputting was blocked. The duration of each pause was 30 sec.

Due to their not being familiar with the task, most subjects required additional verbal instruction at the beginning of the test trial with *ge-*. No additional instructions were necessary for the critical affixes.

Results

Table 1 summarises the results. Columns 2 and 3 in the table list the number of types V_N and the category-conditioned degree of productivity $\boldsymbol{P}$ based on the Eindhoven corpus (Uit den Boogaart, 1975). The numbers of different types produced in the experiment are listed in column 4. What we are especially interested in is how many of these words are "new".

TABLE 1
Observed Number of Types V_N and Category-conditioned Degree of Productivity P as Calculated on the Basis of the Eindhoven Corpus, the Number of Different Types *exp* Produced in the Experiment, the Absolute Number of Neologisms with Respect to the INL Corpus (*n*-INL), the Total Number of Types Attested in the INL Corpus (*t*-INL) and the Number of Types not Listed in the van Dale Dictionary (*n*-vDale)

Affix	V_N	P	*exp*	*n-INL*	*t-INL*	*n-vDale*
-er	299	0.076	447	157	1342	83
-ster	30	0.231	323	255	163	177
-heid	466	0.114	536	130	2399	81
-te	39	0.013	117	71	66	44
her-	24	0.072	257	215	57	182
ver-	355	0.023	378	102	586	27

Unfortunately, it is difficult to define the property of novelty in a rigorous and objective way. I have adopted two practical, if not perfect, strategies. The first uses the Dutch INL corpus (42,000,000 word tokens) as a reference frame. Words not occurring in this corpus are counted as "new". Although the use of this fairly large corpus as a reference frame admittedly involves a simplification, it does provide some insight into which formations are in current use and which are used so infrequently that they are at the very least likely to be processed by rule rather than rote. The second strategy is to use a reliable dictionary as a reference frame. In the present case, the comprehensive van Dale dictionary (Geerts & Heestermans, 1984) has been chosen. Again, this strategy for deciding what is novel is far from perfect. A commercially attractive dictionary cannot attempt an exhaustive listing of all fully regular complex words in a given affix. Such a procedure would lead to an enormous increase in the number of dictionary entries, many of which are totally predictable and hence superfluous. Dictionaries give an indication of the range of use of affixes at best. Although neither strategy is reliable by itself, we may have some confidence when the two strategies for determining novel forms converge. Hence the numbers of types produced in the experiment that do not appear in the INL corpus are listed in column 5 of Table 1 and the numbers of types that are not mentioned in the van Dale dictionary in column 7. Column 6 gives the total number of types that appear in the INL corpus.

What we find is that for each of the affix pairs, the P-based predictions are borne out. The subjects coined some 100 more neologisms in *-ster* than in *-er*, some 50 more in *-heid* than in *-te*, and at least 100 more in *her-* than

in *ver-*. For each of the affix pairs, the difference in the number of neologisms is significant for both the corpus-based and the dictionary-based counts ($P < 0.001$, binomial tests). None of the neologisms in *-ster* is in any way ill-formed. Some of the verbs with *her-*, however, are semantically somewhat odd. Excluding these formations, we are still left with 162 words that are new with respect to the INL corpus, and 133 that are new with respect to the van Dale dictionary. Finally, the majority of the 27 formations with *ver-* that are not registered in the dictionary are semantically well-formed, only one or two being semantically somewhat awkward.

Figure 2 shows how subjects perform the experiment through time for the suffix *-heid*. (For the other affixes studied here, the pattern of results is essentially the same.) The top left-hand panel shows the number of subjects typing in a response during the successive "trials" of the experiment. While all subjects responded with at least 10 words, only one subject typed in 53 words. Apparently, there are substantial differences in the rate at which subjects are able to meet the requirements of the task. The top right-hand panel summarises the frequency in the experiment of the words the subjects produced. The words appearing in the first trials are words that many subjects come up with. As the experiment proceeds, more and more words are produced that are unique in the experiment ($r_s = -0.232$, $P < 0.001$). The next panel shows the number of neologisms (defined with respect to the INL corpus) produced during the successive trials of the experiment. Neologisms appear from the very first trial. The number of neologisms increases slightly up to trial 20, after which their number decreases rapidly, along with the number of subjects responding. The percentage of neologisms on the total of non-erroneous trials increases from an initial 10% to some 20% at trial 28, after which we again observe a decrease. Note, however, that the number of subjects responding is halved by trial 30. In fact, the two subjects with more than 37 responses produce 30 words during the last 16 trials, of which 7 are neologisms (23%). This suggests that if the same number of subjects had responded up to the last trial, the downward curvature would have been absent. We may conclude that the main trend is an increase in the number of neologisms through time. This ties in with the opposite pattern we observed for the experimental word frequencies. Initially, subjects tend to come up with the same types. As the experiment continues, they branch out in different directions through the lexicon, producing less common words and increasing numbers of neologisms.

Note that from the very first trial onwards, neologisms are produced. This shows that Strategy 3 is exploited from the start. Hence the theoretical possibility that subjects first scan their lexicons for existing words in *-heid*, and only then start to coin new words, can be ruled out. Given that subjects use Strategy 3, we may expect words with high-frequency base

words to appear in the earlier trials, and words with lower-frequency base words in the later trials. The lower right-hand panel in Fig. 2 shows this to be the case: trial number and base word frequency are negatively correlated ($r_s = -0.1453$, $P < 0.001$). The correlation, however, is very weak. Inspection of the sequence of trials for individual subjects shows that semantic relations such as antonymy and synonymy are heavily exploited. These relations substantially weaken the base frequency effect: in the

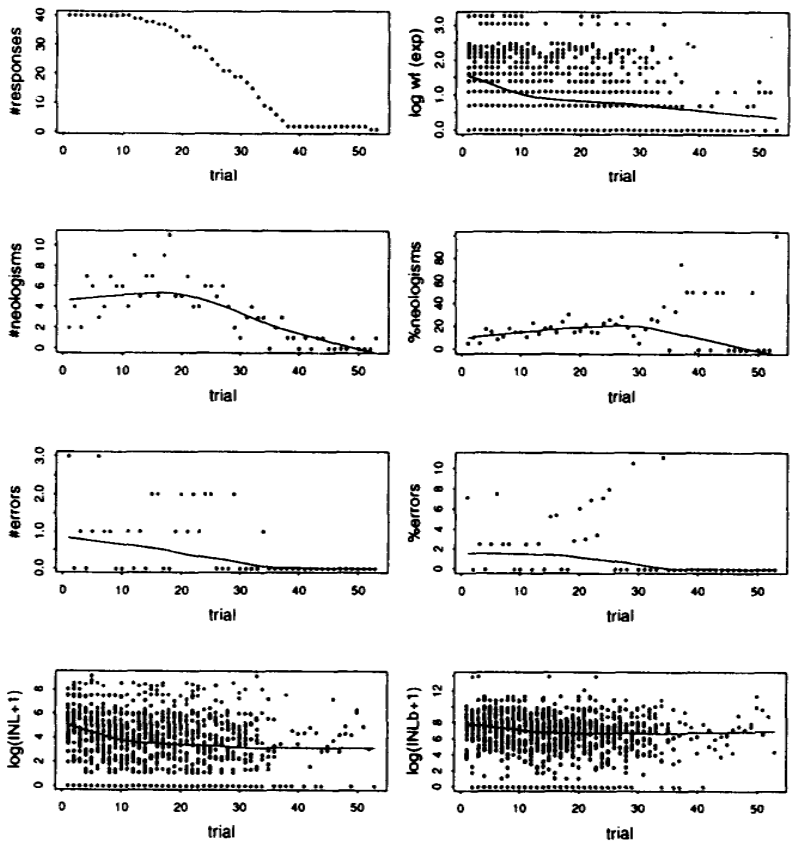

FIG. 2 Trial statistics for the suffix *-heid*. The horizontal axis represents the successive trials in the experimental frequency (wf(exp)) of the responses, the number of neologisms (defined with respect to the INL corpus) produced at trial t, the percentage of neologisms, the number of errors and the corresponding percentages, the frequency of the words in the INL corpus (INL), and the frequency in the INL corpus of the corresponding base words (INLb).

responses of our first subject, we find sequences such as *grootheid* ("largeness"), with a base frequency of 51,668, followed by *kleinheid* ("smallness"), with a base frequency of 21,534; *zachtheid* ("softness"), with a base frequency of 8276, followed by *ruwheid* ("coarseness"), with a base frequency of 1624; and *moeheid* ("tiredness"), with a base frequency of 2547, followed by *fitheid* ("fitness"), with a base frequency of 146. In fact, *fitheid* is a neologism with respect to both corpus and dictionary, illustrating how subjects happen to chance on base words for which the abstract noun in *-heid* does not exist. The semantic heuristics used in Strategy 3 similarly explain why trial number and word frequency (in the INL corpus) are only weakly correlated ($r_s = -0.1936$, $P < 0.001$).

Next let us consider Table 2, which lists the errors (pseudo-affixed words, ungrammatical formations, inflectional intrusions, etc.) produced in the experiment. The proportion of errors for *-er* is significantly higher than that for *-ster* ($Z = 3.68$, $P < 0.001$); the same is true for *-te* compared with *-heid* ($Z = 17.33$, $P \ll 0.001$). The proportions of errors for *her-* and *ver-* do not differ significantly ($Z = 0.71$, $P > 0.10$). The presence of pseudo-affixed words such as *geheid* ("surely"), *boete* ("fine"), *moeder* ("mother"), *lijster* ("thrush"), *herten* ("deer") and *verbaal* ("verbal") show that form-based scans of the lexicon (Strategy 1) are initiated. Some of the errors reveal the effects of a rhyme-based search (*af+scheid*, "departure"). Inflectional intrusions involving the homonymic past-tense marker *-te* account for 43% of the errors for derivational *-te*. Similarly, the comparative suffix *-er* accounts for 43% of the errors for derivational *-er*. The two panels on the third row of Fig. 2 summarise for the suffix *-heid* where the errors appear in the sequence of trials. What we observed is an effect of learning: later in the experiment, fewer errors appear. From trial 35 onwards, the subjects who are still responding perform error-free, suggesting that they no longer use Strategy 1 and exclusively exploit

TABLE 2
Absolute Number of Errors and Error Percentages for the Six Affixes (Type-based)

Affix	*Correct*	*Errors*	*Errors/(Correct + Errors)*
-er	447	114	0.203
-ster	323	40	0.108
-heid	536	18	0.032
-te	117	146	0.555
her-	257	16	0.059
ver-	378	31	0.076

TABLE 3
Lexical Statistics of the Generated Word-frequency Distributions: The Number of Tokens N, the Number of types V_N, the Number of Hapaxes or Unique Types $V_N(1)$, the Standard Deviation of $V_N(1)$, the Growth Rate P and the Population Number of Types $\hat{V}$

Affix	N	V_N	$V_N(1)$	$\hat{\sigma}(V_N(1))$	P	$\hat{V}$	$\chi^2_{(8)}$	q
-er	1064	447	258	14.14	0.242	2023	8.11	0.423
-ster	838	323	197	12.56	0.235	2905	5.21	0.735
-heid	986	536	348	16.37	0.353	2176	26.45	0.001
-te	509	117	50	6.03	0.098	254	11.14	0.194
her-	791	257	149	10.99	0.188	3683	5.19	0.737
ver-	1224	378	158	10.06	0.129	678	2.71	0.951

Growth rate and standard deviations were estimated on the basis of the generalised inverse Gauss–Poisson model, the goodness-of-fit statistics of which are given in the last two columns.

Strategy 3. The absence of any curvature for these trials in the scatterplots of word frequency (INL) and base frequency (INL) support this conclusion: subjects associatively scan their lexicons for the appropriate base words.

Table 3 summarises the main statistical properties of the word-frequency distributions generated in the experiment. The greatest number of tokens was generated for the prefix *ver-*, the greatest number of types for *-heid*. The latter suffix also appears with the highest number of hapaxes and the highest growth rate. Its counterpart *-te* appears with only 50 hapaxes and has the lowest growth rate. It is interesting to observe that the lowest number of types in the population as estimated on the basis of Sichel's generalised inverse Gauss–Poisson "law" (Sichel, 1986; see also Baayen, 1993a; Chitashvili & Baayen, 1993), are calculated for *ver-* and *-te*, the affixes with the lowest category-conditioned degrees of productivity in both corpus (Table 1) and experiment (Table 2).[5]

The sampling time for each of the six affixes studied here was exactly the same. This allows us to interpret the growth rates of the six word-frequency

[5]In the light of the extremely small sample sizes, the absolute values of the estimated population vocabulary sizes can be interpreted as rough indicators of their general order of magnitude only, small changes in $V_N(m)$ for small m having the potential of substantially reducing or increasing V. In the case of *-ster* and *her-*, this may lead to an overestimation of the number of types. On the other hand, *her-* and *-ster*, the two affixes for which the estimated number of types is extremely high, are semantically completely transparent, allowing subjects to produce new words without hesitation. Possibly, the effect of semantic transparency on productivity is measurable in terms of $\hat{V}$.

distributions obtained in the experiment directly as measures for the degree of productivity without further conditioning, each distribution being an independent sample in its own right and not a subset of some larger sample (corpus or text). Comparing the (experimental) degrees of productivity for each of the three pairs of affixes, we find, using the test statistic:

$$Z = \frac{V_N(1, e)/N_e - V_N(1, f)/N_f}{\sqrt{\sigma(V_N(1, e))^2/N_e^2 + \sigma(V_N(1, f))^2/N_f^2}} \quad (4)$$

that *-er* and *-ster* have the same growth rate ($Z = 0.37$, $P > 0.20$) and that *-heid* and *her-* have significantly higher growth rates than *-te* ($Z = 12.49$, $P < 0.001$) and *ver-* ($Z = 3.67$, $P < 0.001$), respectively.

Note that of all the affixes considered here, it is the one with the highest category-conditioned degree of productivity in Table 3 (*-heid*) that appears with the lowest error rate. Conversely, the affix with the lowest experimental category-conditioned degree of productivity (*-te*) appears with the highest error rate.

Discussion

For each of the pairs studied here, the category-conditioned degree of productivity correctly predicted for which affix the largest number of neologisms was generated. It is worthwhile to consider the results obtained for each of the affix pairs in some more detail, since this will allow us to isolate some of the factors determining the productivity of the affixes involved. First, consider the rival affixes *-te* and *-heid*. We have seen that subjects create significantly more neologisms in *-heid* than in *-te*. In addition, the number of different types in *-te* realised in the experiment ($n = 117$) is only one-quarter of the number of types in *-heid* ($n = 536$). The experimental degree of productivity of *-heid* (0.353) is very much larger than that of *-te* (0.098), and the estimated numbers of types that could in principle have been generated in an infinitely long experiment (2176 for *-heid*, 254 for *-te*) are likewise substantially different. Taken together, we have converging evidence that *-heid* is a fully productive affix and that *-te* is marginally productive at best. Perhaps the most important factors underlying the unproductiveness of *-te* are (1) that it attaches to monomorphemic adjectives only, and (2) that there is a largely synonymous affix (*-heid*) that attaches to both simplex and complex adjectives. With this in mind, consider again the strategies subjects may employ when asked to type words with *-te* into the computer. On the one hand, they may use Strategy 1 and search their lexicons for existing formations. Since the frequency distribution of *-te* is dominated by high-frequency types, this is a useful

strategy. On the other hand, they may attempt to use Strategy 3, scanning the lexicon for appropriate base words. This strategy will be relatively unsuccessful, however, given the small number of simplex adjectives available in Dutch that may serve as base words and the greater naturalness of the rival formations in *-heid*. Hence subjects may be expected to depend rather heavily on a form-based scan of the lexicon. With a form-based scan, the risk of intruding forms ending in the string *te* without containing the required de-adjectival suffix *-te* is high. This is what we observed in the experiment: under time pressure to produce as many forms as possible for a suffix with only some 40 words in daily use in the language, the experimental subjects produced a substantial number of errors. This pattern is completely reversed for *-heid*. Its frequency distribution is dominated by low-frequency types. Although subjects may opt for scanning the lexicon for the relatively small number of well-established "existing" nouns in *-heid*, it is far more productive to apply the word formation rule to the many complex adjectives to which *-heid* attaches so easily. Although the large number of neologisms obtained shows that subjects have indeed made use of productive word formation, some care is required with respect to the interpretation of the low error rate of *-heid*. In general, a higher dependence on productive word formation will lead to fewer form-based errors, but in this case it should be noted that such errors are extremely unlikely to occur for the simple reason that, apart from *geheid* ("surely") and a scattering of words ending in the string *scheid*, no homonymic word-final strings exist in the language. This, of course, may well be one of the reasons that *-heid* is more productive than *-te*.[6]

Next consider the suffixes *-er* and *-ster*. As the unmarked member of the pair, *-er* is used far more intensively in the language than *-ster*. Nevertheless, subjects are able to coin a substantial number of words in *-ster*, be it not as many as in *-er*. Evidently, *-ster* is a fully productive suffix of Dutch, as predicted by $\boldsymbol{P}$, even though it is used sparingly. There are two questions that invite further comments. First, why do we find an error rate of 1:5 for *-er*? And second, why is the category-conditioned degree of productivity $\boldsymbol{P}$ of *-ster* in the Eindhoven corpus higher than that of *-er*, while its degree of productivity as measured in the experiment does not differ significantly from that of its unmarked counterpart? As will be argued below, task-specific requirements may underlie the high $\boldsymbol{P}$ value for *-ster* in the experiment. Turning to the question why *-er* appears with such a high error rate, the following observations may be relevant. First, the frequency distribution of *-er* is characterised by a rather large number of high-

[6]For the possible effect of large numbers of pseudo-affixed words on productivity, the reader is referred to Baayen (1993b).

frequency formations. This suggests that a memory scan for existing formations (Strategy 1) may be relatively successful. Second, unlike *-ster*, *-er* has a homonymic inflectional rival, comparative *-er*. A fairly large number of the erroneous responses for agentive *-er* are comparative formations. Their appearance is probably due to the interference of existing comparative formations during scanning operations for existing words with agentive *-er*. Third, *-er* is known to have a wide range of semantic readings ranging from personal agent (*zender*, "sender") through impersonal agent (*zender*, "radio station") to instrument (*zender*, "transmitter"), and from object names (*instapper*, "a shoe without shoe laces") through event names (*misser*, "failure") to causative names (*giller*, "what makes one scream") (cf. Booij, 1986). Even though the personal agent reading is the most dominant one, the procedural knowledge for coining nouns in *-er* may well require more specific instructions from the conceptualiser than were available to our experimental subjects. In other words, Strategy 3 may have been at a disadvantage compared with *-ster*, where the semantics are straightforward. Considered jointly, these observations strongly suggest that our subjects have been exploiting their knowledge of the forms of existing nominalisations in *-er* intensively, with a high error rate as a result.

Finally, consider the prefixes *her-* and *ver-*. The latter prefix is characterised by a complex lexical conceptual structure that can be realised in a variety of ways (Lieber & Baayen, 1994), whereas the former prefix has a simple adverbial reading only ("again"). This suggests that the processing costs of creating neologisms in *her-* should be less than those for *ver-*. Hence a greater number of neologisms in *her-* is expected. This expectation is borne out by the present experiment. Conversely, since the frequency distribution of *ver-* is dominated by high-frequency types, scanning lexical memory for existing words with *ver-* will be a particularly productive strategy for meeting the requirements of the task. Not surprisingly, subjects produce many more well-established formations in *ver-* than in *her-*. In the absence of homophonic prefixes, *her-* and *ver-* give rise to approximately the same low error rate. Note that our subjects produced more words in *ver-* than in *her-*, where most words in *ver-* are "existing" words of the language, while the majority of the *her-* words are "new". This pattern of results suggests that the on-line creation of neologisms requires more time than the retrieval of known words from the lexicon. This may also explain why the number of types produced for *-er* ($n = 447$) exceeds that in *-ster* ($n = 323$).

Summing up, we can conclude that this experiment shows that the ***P*** statistic can be reliably used to trace whether a word formation rule is available *in principle* for the coining of new complex words. For affixes such as *her-* and *-ster* that are productive *in principle* in the sense of

Schultink (1961), the high value of P is a signal that many new formations can be formed, if required. Whereas *ver-* and *-er* have already attached to a large number of the available base words, *-ster* and *her-* have attached to a small number of these base words only. Using Strategy 3 for *her-* and *-ster*, subjects will chance upon more base words that happen not to have the corresponding complex word than in the case of *ver-* and *-er*. In other words, if an absolute distinction between productive and unproductive rules is to be made, P is to be preferred to P^*.

It is less clear whether the experimental results can be used to check the validity of the interpretation of P as the likelihood of coining a new word at the level of lemma retrieval and phonological encoding. First, in normal language use, on the basis of which P is calculated, words are often used repeatedly. In the experiment, however, subjects are instructed to avoid producing a word more than once. This difference, which is especially important for the statistical model underlying the P statistic, may underlie the divergence between the P values of *-er* and *-ster* measured in the corpus and those measured in the experiment. A related problem is that the form-based scanning strategy is task-specific, whereas the strategy in which the lexicon is scanned for the appropriate base words is a better approximation of normal lemma retrieval and phonological encoding. As we have seen, the form-based strategy introduces a large amount of erroneous responses due to form-similarity. More precise results may be expected in tasks where the use of this strategy is ruled out. Second, if a measure such as P is to be used for gauging concept-driven productivity in production, it should not be calculated on the basis of the words with a given affix only. Words with (roughly) synonymous affixes and monomorphemic words belonging to the same target semantic category should also be included.

GENERAL DISCUSSION

I have argued on the basis of lexical statistics that even though the role of procedural lexical knowledge in speech production may be marginal token-wise, the rate at which new words are generated is high enough for morphological productivity to be a relevant concept for psycholinguistic theories of speech production. The relevance of this notion was confirmed by a production experiment which showed that subjects are able to coin hundreds of well-formed new morphologically complex words for those affixes which are semantically transparent and productive. The pattern of results obtained contrasts with the creative use of derivational morphology by three Italian jargonaphasics reported by Semenza et al. (1990). These patients were found to produce large numbers of derivationally complex neologisms which were morphologically well-formed according to the rules of Italian. The majority of these neologisms, however, exploited word

formation patterns that are judged to be unproductive by native speakers of the language.

Can this be taken to imply that these patients are forced by their language deficit to exploit the least constrained and perhaps basic means for creating new words, namely analogical modelling? To my mind, such a conclusion may be premature. Paradoxically, it may well be that the observed intensive use of unproductive word formation patterns is a direct consequence of their being unproductive. To see this, first recall that unproductive word formation patterns are characterised by high frequencies of use. In fact, it is not uncommon for a less productive or unproductive affix to show up with more tokens in a corpus than its productive counterpart. For instance, English *-ness*, which is highly productive, is realised in 22,856 word tokens in the Cobuild corpus. Of the 1607 different types in *-ness*, nearly half are hapaxes ($V_N(1) = 749$). In contrast, the morphological rival *-ity* appears in 45,861 tokens, distributed over 734 types. There are only 218 hapaxes. The average frequency of a formation in *-ness* equals 14.2, whereas for *-ity* the mean frequency is roughly four times as high (62.5). For *-ness* the median is 2, for *-ity* it is 5.[7] Assuming that high frequencies of use facilitate phonological encoding and subsequent articulatory processes, unproductive affixes will have an advantage with respect to their productive counterparts in the later stages of the production process. Second, it should be noted that many unproductive word formation patterns have productive, more or less synonymous counterparts. When the conceptualiser submits a pre-verbal message to the formulator, the process of grammatical encoding is at a certain point faced with the choice between two or more alternative realisations. Assuming that this choice is influenced by the semantic transparency of the affix, the more productive alternative will be selected in unimpaired speech. For the aphasics studied by Semenza et al. (1990), the semantic system is not available. In the absence of semantic transparency to guide the choice, some other criterion, such as the activation level of the form representations of the affixes, has to be found. Here the productive affix is at a disadvantage, however, which immediately leads to the observed favouring of the unproductive rival form in aphasic speech. Although admittedly speculative, this explanation, which hinges on the strong links between productivity, transparency and frequency, allows at least a glimpse into why the preference for the productive member of a set of rival affixes in

[7]These counts are based on a more precise count of words in *-ness* and *-ity* than was available to me when I studied the productivity of English derivation with R. Lieber. Not surprisingly, the more accurate counts bring out more clearly the marked difference in productivity between *-ness* and *-ity* than the counts presented in Baayen and Lieber (1991).

normal speech is reversed in the semantically anomalous speech of jargonaphasics.

Manuscript received 18 February 1993
Revised manuscript received 16 November 1993

REFERENCES

Anshen, F., & Aronoff, M. (1988). Producing morphologically complex words. *Linguistics, 26*, 641–655.

Baayen, R.H. (1989). *A corpus-based approach to morphological productivity: Statistical analysis and psycholinguistic interpretation*. Dissertation, Free University, Amsterdam.

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. (1993a). Statistical models for word frequency distributions: A linguistic evaluation. *Computers and the Humanities, 26*, 347–363.

Baayen, R.H. (1993b). 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, 2*, 801–843.

Baayen, R.H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bolinger, D.L. (1948). On defining the morpheme. In D.L. Bolinger (Ed.), *Forms of English: Accent, morpheme, order*, pp. 183–189. Cambridge, MA: Harvard University Press.

Booij, G.E. (1986). Form and meaning in morphology: The case of Dutch "agent nouns". *Linguistics, 24*, 503–517.

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

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.

Chitashvili, R.J., & Baayen, R.H. (1993). Word frequency distributions. In L. Hřebíček & G. Altman (Eds), *Quantitative text analysis*, pp. 54–135. Trier: Wissenschaftlicher Verlag Trier.

Geerts, G., & Heestermans, H. (1984). *Van Dale Groot Woordenboek der Nederlandse Taal*. Utrecht: Van Dale Lexicografie.

Geerts, G., Haeseryn, W., de Rooij, J., & Van den Toorn, M.C. (Eds) (1984). *Algemene Nederlandse Spraakkunst*. Groningen: Wolters-Nordhoff.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lieber, R. & Baayen, R.H. (1994). Dutch verbal prefixation: Lexical conceptual structure, argument structure and productivity. In G.E. Booij & J. Van Marle (Eds), *Yeaook of morphology 1993*, pp. 51–78. Dordrecht: Kluwer.

Maclay, H., & Osgood, C.E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*, 19–44.

Muller, Ch. (1979). *Langue Française et linguistique quantitative*. Genève: Slatkine.

Renouf, A., & Baayen, R.H. (1994). *Chronicling The* Times: *Lexical innovations in a British newspaper*. Unpublished manuscript, Max-Planck Institute for Psycholinguistics, Nijmegen.

Schultink, H. (1961). Productiviteit als Morfologisch Fenomeen. *Forum de Letteren*, *2*, 110–125.

Semenza, C., Butterworth, B., Panzeri, M., & Ferreri, T. (1990). Word formation: New evidence from aphasia. *Neuropsychologia*, *28*, 499–502.

Sichel, H.A. (1986). Word frequency distributions and type-token characteristics. *Math. Scientist*, *11*, 45–72.

Uit den Boogaart, P.C. (Ed.) (1975). *Woordfrekwenties in Geschreven en Gesproken Nederlands*. Utrecht: Oosthoek, Scheltema and Holkema.

Van Marle, J. (1985). *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris.

Van Marle, J. (1992). The relationship between morphological productivity and frequency: A comment on Baayen's performance-oriented conception of morphological productivity. In G.E. Booij & J. Van Marle (Eds), *Yearbook of morphology 1991*, pp. 151–163. Dordrecht: Kluwer.

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