Productivity and English derivation: a corpus-based study*

HARALD BAAYEN and ROCHELLE LIEBER

Abstract

The notion of productivity is one which is central to the study of morphology. It is a notion about which linguists frequently have intuitions. But it is a notion which still remains somewhat problematic in the literature on generative morphology some 15 years after Aronoff raised the issue in his (1976) monograph. In this paper we will review some of the definitions and measures of productivity discussed in the generative and pregenerative literature. We will adopt the definition of productivity suggested by Schultink (1961) and propose a number of statistical measures of productivity whose results, when applied to a fixed corpus, accord nicely with our intuitive estimates of productivity, and which shed light on the quantitative weight of linguistic restrictions on word-formation rules. Part of our purpose here is also a very simple one: to make available a substantial set of empirical data concerning the productivity of some of the major derivational affixes of English.

In this paper we propose a measure of productivity in morphology which is based on the definition of productivity in Schultink (1961). We argue that a measure of productivity based on the token frequencies of types, specifically on the number of hapax legomena for a given affix in a corpus, comes very close to according with our intuitions about productivity. We illustrate this result by applying our measure to a substantial body of empirical data from English derivational morphology. Our aim is not merely to develop a quantitative measure and to see how it accords with the data, but also to provide a measure of productivity that would be of use in morphological theory. Specifically, having such a measure of productivity would be of use in delimiting the set of data which a theory of word formation should be accountable to. Presumably, morphological theory should account only for processes of word formation which are

Linguistics 29 (1991), 801–843 0024–3949/91/0029–0801 $2.00

productive; processes of word formation that are no longer productive or putative morphological processes that never have been productive at all are of little or no interest to morphological theory.[1] In section 1 we review past attempts at quantifying productivity. Section 2 develops our proposed measure, and section 3 applies this measure to a selection of English derivational affixes. In section 4 we outline a complementary technique that can be used to quantify productivity.

1. Background

Aronoff (1976) represents the first attempt in the literature on generative morphology to formalize the notion of productivity. He points out that the notion, although widely used, is often left rather vague. Simple attempts to measure productivity, for example, counting up numbers of words with particular affixes (the more words with some affix, the more productive the affix), fail to coincide with our intuitive sense of productivity. One problem with a simple counting method is that

> ... it doesn't take into account the fact that there are morphological restrictions on the sorts of words one may use as the base of certain WFRs [word-formation rules; HB and RL]. Thus *#ment* and *+ion* both form nouns from verbs (*detachment, inversion*), but the latter is restricted to latinate verbs. There is a simple way to take such restrictions into account: we count up the number of words which we feel could occur as the output of a given WFR (which we can do by counting the number of possible bases for the rule), count up the number of actually occurring words formed by that rule, take a ratio of the two, and compare this with the same ratio for another WFR. In fact, by this method we could arrive at a simple index of productivity for every WFR: the ratio of possible to actual words (1976: 36).

There are a number of problems with Aronoff's suggested method of measuring productivity. The first is a very general one that has to do with the notion of counting 'actual' words. Words are not 'actual' or 'existing' in any objective sense. A list of actual words always involves a subjective element: they are words compiled in the mental lexicon of some individual or other, or in a dictionary produced by certain lexicographers, or in a fixed corpus constructed in some more or less arbitrary way. The 'actual' words of any of these sources will very likely coincide in large part with the actual words of the others, but perhaps never in totality. In other words, the notion of 'actual' word is to some extent a fiction, although a convenient and necessary one if we are to count words and thereby measure productivity. In order to make use of Aronoff's (or for that matter anyone's) measure of productivity at all, then, we must first agree on a reasonable list of 'actual' words.

This problem is, of course, not insurmountable; some large and varied but fixed sample could potentially give us a representative-enough slice of a language on which to base our counts. This is the choice we will make here; we will make use of a large English lexical database of 18 million word forms from the Dutch Centre for Lexical Information in Nijmegen, CELEX, version E1.0. This lexical database, henceforth refered to as the 'Celex database', has been compiled on the basis of the corpus of the Cobuild project of the University of Birmingham (Renouf 1987). The corpus is taken from both written and spoken language (75% written, 25% spoken), from the following categories: 'broadly general, rather than technical, language; current usage, from 1960, and preferably very recent; "naturally occurring" text, not drama; prose, including fiction and excluding poetry; adult language, 16 years or over; "standard English", no regional dialects; predominantly British English, with some American and other varieties' (Renouf 1987: 2).[2]

Such a corpus is superior to, for example, the list of words in a dictionary. First, it offers information about the frequency of words (information which will be of importance below). Second, it contains words of the sort that dictionaries typically do not list (such as words formed with highly productive affixes like *-ness*). Third, dictionaries may list words which are not used in actual speech. For instance, as pointed out by Anshen and Aronoff (1988: 645), even though Walker (1936) lists 23 words in *-ivity* and 27 words in *-ibleness*, only the former words are attested in the Kučera and Francis (1967) corpus.

Assuming that we have some reasonable way of characterizing the notion of 'actual' word, then, Aronoff's suggested index of productivity might be formalized as (1), as suggested in Baayen (1989):

(1) $$I=\frac{V}{S},$$

where I = index of productivity, V = the number of types, and S = the number of types the WFR in question could have given rise to.[3] (We use here and below the familiar distinction between type — the number of different forms occurring with a particular affix — and token — particular instances of a given type.) The index of productivity (1) is not without problems, however. Note, first, that comparing the number of types that particular affixes give rise to often produces counterintuitive results. For example, in a study of Dutch word formation, Baayen (1989) points out that the noun-forming suffixes *-te* (*zwak-te* 'weakness') and *-sel* (*voed-sel* 'food') exhibit comparable numbers of types, yet native speakers of Dutch intuitively feel *-sel* to be productive, and *-te* not or hardly to be productive. Similarly, for English, the noun-forming suffixes *-ness* (*happiness*) and *-ity* (*purity*) show respectively 497 and 405 types in the Celex corpus, yet the former is felt to be productive to a higher degree than these numbers of types suggest. Even more counterintuitive are the type figures for the English verb-forming prefixes *de-* (*debug*) and *en-N* (*enthrone*). *En-N* shows 40 types, *de-* only 32 types, yet the latter is intuitively felt to be much more productive than the former.

A further problem with (1), also discussed in some detail in Baayen (1989), concerns the figure *S*; it is not necessarily clear how to count up the number of types which could POTENTIALLY be created with a given affix. For example, suffix Z may be productive with words formed with suffix X. If suffix X is itself very productive, there may be no reasonable way of estimating how many forms there are in suffix X; dictionaries, after all, would not necessarily contain such productively derived forms. Baayen (1989: 30) in fact points out, '... the index of productivity vanishes for productive word formation rules. ... The index is, in fact, applicable to unproductive word formation rules only, and is perhaps better named an index of unproductivity.' That is, as more and more actual words are taken into account by considering a sequence of corpora of increasing size, the index *I* will approach unity for unproductive word-formation processes ($V \rightarrow S$), but will remain zero for productive rules where *S* is, at least in theory, (enumerably) infinite.

In subsequent research, Aronoff also takes into account the token frequencies of derived formations. In Aronoff (1982) he calls attention to the fact that words of the form *Xivity* have a higher mean frequency than words of the form *Xiveness*. On the basis of the relevant types listed in Walker (1936) and their token frequencies in the (1,000,000) Kučera and Francis (1967) word list, he arrives at a mean frequency of 9.565 tokens/type for the 23 types *Xivity* and a mean frequency of 0.641 for the 103 types in *Xiveness*. This pattern repeats itself for the 18,000,000 Celex database, as shown in Table 1, where only those types which occur in the database are taken into account.[4]

Since the frequency distribution of the raw data is highly skewed, with the bulk of the types having frequencies which are much lower than the mean frequency, rather than frequencies clustered around the mean frequency, significance testing on the basis of $\hat{\mu}$ and $\hat{\sigma}$ is not possible. However, when we consider the logarithms of the token frequencies rather

Table 1. *Mean frequencies of* Xiveness *and* Xivity *in the Celex database*

	V	N	$\hat{\mu}=N/V$	$\hat{\sigma}$	$\hat{\mu}_{\log}$	$\hat{\sigma}_{\log}$
Xivity	18	3692	205.11	630.31	3.40	1.76
Xiveness	27	465	17.22	32.57	1.75	1.37

Key
V: number of types
N: number of tokens
$\hat{\mu}$: sample mean token frequency
$\hat{\sigma}$: sample standard deviation
$\hat{\mu}_{\log}$ and $\hat{\sigma}_{\log}$: sample mean and standard deviation under the lognormal hypothesis

than the raw frequencies themselves, the highly skewed frequency distributions are transformed into ones which are approximately normal. For such so-called lognormal distributions we can test whether the mean token frequency of formations in *Xivity* is significantly higher than that of formations in *Xiveness*. On the basis of the respective values of $\hat{\mu}_{\log}$ and $\hat{\sigma}_{\log}$ we find that the null hypothesis that the mean token frequency of *Xivity* is not larger than that of *Xiveness* can be rejected ($Z=3.34$, $p<0.001$). In his (1982) article, Aronoff interprets this finding in the light of the fact that the formations in *-ity* 'are more likely to be lexicalized and assigned special meanings. We now see that this lexicalization is reflected in frequency, for semantic complexity and frequency go hand in hand.'[5]

A more detailed interpretation of the relation between token frequency and productivity is presented in Anshen and Aronoff (1988). Their central idea is

> (1) that people do, in fact, store certain complex morphological items in their mental lexicons while they construct others as needed; (2) that in producing sentences, speakers simultaneously attempt to find a needed lexical item and to build it by rule from a related form (1988: 642).

On the basis of a production test in which subjects were required to compile a list of words in *-ibleness*, *-ibility*, *-iveness*, *-ivity*, *-ionary*, and *-ional*, a test which showed that subjects are somewhat more likely to coin nonce words in *-ibleness* and *-iveness* than nonce formations in *-ibility* and *-ivity*, Anshen and Aronoff argue that *-ity* forms are stored in the mental lexicon, while the forms in *-ness* are not stored at all but constructed by rule as needed. Also, 'if speakers construct *-ness* forms freely, while picking *-ity* forms from a defined set, it is reasonable to predict a wider scatter (that is, a lower type–token ratio) for *-ness* words than for *-ity* words.'

This is what they found in their experimental data, and what can be observed for the frequency distributions of *-ness* and *ity* in corpora. Note, however, that the observation of a lower type–token ratio for *-ness* is logically independent of the claim that words in *-ness* are not stored in the mental lexicon, since a low type–token ratio for *-ness* can be the result of the simple fact that that these words are being sampled from a larger population of formations. As the number of types in the population increases, the chance of sampling some particular type more than once decreases. Consequently, the low type–token ratio for *-ness* may simply be the result of the fact that the number of possible words in *-ness* is very large. Moreover, it is rather counterintuitive to claim that no formations in *-ness* are stored, since there is some overlap in the token-frequency ranges of formations in, for instance, *-ivity* and *-iveness*, as shown in Table 2, and since it is unlikely that concepts like *forgiveness* or *effectiveness* are reinvented for each successive instance of use. Summing up, we find the hypothesis that productive formations are not listed, in contrast to less productive or unproductive ones, too simplistic. Consequently, their argument that the existing (higher-frequency) formations in *Xivity* block the corresponding formations in *Xiveness* (for which [token] frequency is judged to be irrelevant), while at the same time the rule-generated formations in *-iveness* block 'access to and thus the existence of a lexically based *-ivity* form' (1988: 653), cannot be correct. In other words, Anshen and Aronoff's (1988) attempt to find frequency-based support for Aronoff's (1976) blocking analysis fails, both (1) logically since, as pointed out by van Marle (1985), words cannot at the same time block and be blocked, when blocking is interpreted in terms of high-frequency existing words preempting the coining of novel, 'zero-frequency' formations;[6] and (2) empirically, since there is no a priori reason to suppose that high-frequency formations in *-ness* are not stored. In what

Table 2. *Token frequencies of the five most frequent types in* -ivity *and* -iveness *in the Celex database*

-ness	Frequency	-ivity	Frequency
effectiveness	141	activity	2785
forgiveness	91	productivity	331
permissiveness	59	sensitivity	150
agressiveness	35	relativity	65
destructiveness	21	passivity	55

follows we will make the simple assumption that derived formations are more likely to be stored as their token frequency increases, whatever the productivity of the underlying word-formation process may be.

The main focus of Anshen and Aronoff (1988), however, is on the relation between base word frequency and the frequency of the corresponding derived word. Their basic observation concerns the fact that, for example, the irregular plural *children* has a higher frequency than *child* (11,656 against 7,619 in the Celex database). They argue that the irregular plurals block the formation of the corresponding regular plurals. However, apart from such extreme cases, which typically arise in inflection, where the distinction between productive and unproductive is very clear-cut (Scalise 1988), the relation between the frequencies of base and derivative are not particularly relevant to the study of productivity in derivation, where one is confronted with the problem of varying degrees of productivity. To show this, we call attention to a study by Thorndike (1943), who introduced the notion of derivation ratios, the ratio of the number of derived tokens to the number of base tokens, calculated for each type. When such derivation ratios are calculated for productive WFRs, distributions of derivation ratios are obtained that show a wide range of possible shapes, scarcely narrower than the theoretically possible maximum range. Moreover, the distributions obtained for unproductive WFRs fall within the same range. Hence it is impossible to distinguish between productive and unproductive WFRs on the basis of these derivation ratios. In fact, derivation ratios are, at least in part, semantically determined. For instance, *redly* occurs only once compared with 1,972 occurrences of *red* in Thorndike's corpus, a derivation ratio of 0.001 to three decimal digits, and the derivation ratios for *blackly* and *whitely* are 0.002 and 0.000 respectively. In contrast, such adverbs as *recently* and *slowly* show up with derivation ratios well above unity, that is, these adverbs occur more often than their adjectival bases. As pointed out by Thorndike (1943: 34), adverbs are hardly ever coined from adjectives which describe sensory qualities: 'For we often need to state than an object can produce that sensation or has that quality, but relatively seldom need to state that anything is acting in that way.'

Similarly, the derivation ratios for the Dutch derivational suffixes *-te* (unproductive) and *-heid* (productive) do not reflect differences in productivity but, as in the case of English *-ly*, have a bearing on the semantics of the underlying base words (Baayen 1989: 268–274).[7]

Whereas derivation ratios are irrelevant to the issue of productivity, it is profitable to consider the frequency distributions of the base words underlying complex formations. Harwood and Wright (1956) are, to our knowledge, the first to call attention to a striking difference between the frequency distributions of derived words and the corresponding distributions of their bases, namely that the distributions of the derived words are characterized by substantially larger numbers of very low frequency types than the associated base word distributions. They argue that this difference counts as evidence for a theoretical approach to derivation in which users are regarded as having 'a stock of bases and mechanisms for making further words by adding ... word-forming elements ...' (1956: 260). Their aim is to show that statistical study of English word formation can shed light on these mechanisms or, in present-day terminology, on the generative aspect of word-formation rules. They also remark that their data may 'provide ways of measuring the extent to which a suffix is "living"' (Harwood and Wright 1956: 263). Although they are not explicit about what kind of measure might be relevant, their insight can be exploited to yield a measure of productivity. Before we introduce such a measure, however, we first state more precisely what our view on productivity is.

To our minds, the best intuitive definition of productivity is that given by Schultink (1961) (translation from van Marle 1985: 45):

> Onder produktiviteit als morfologisch fenomeen verstaan we dan de voor taalgebruikers bestaande mogelijkheid ... onopzettelijk een in principe niet telbaar aantal nieuwe formaties te vormen.
>
> [By productivity as a morphological phenomenon we understand the possibility for language users to coin, unintentionally, a number of formations which are in principle uncountable ...].

There are two important features of Schultink's definition of productivity. The first has to do with the notion of 'unintentionality'. If a word-formation process is truly productive, new formations using that process will go unnoticed. For unproductive processes a new form may sometimes be coined, but such coinages will always draw attention to themselves: they will be used to shock, to amuse, or to achieve some other intentional effect. The second feature of Schultink's notion of productivity is the idea of countability: truly productive word-formation processes will give rise to in principle infinite numbers of new forms, while unproductive word-formation rules will give rise to a fixed, and therefore countable, number of forms. In this way productivity, as Lyons (1977: 549) puts it, a design feature of the language, is distinguished from creativity, the language user's ability to extend the vocabulary by means of motivated, but unpredictable, principles of abstraction and comparison.

Schultink does, of course, allow that among productive processes of word formation some might be more productive than others. For example, some affixes have phonological, syntactic, or semantic restrictions which prevent them from attaching to bases of certain sorts. The English comparative suffix *-er*, for example, attaches roughly to adjectives of two or fewer syllables, where the second syllable must be weak (*happier*, **directer*, **intelligenter*). The suffix *-able*, which forms adjectives from verbs, attaches only to verbs with an appropriate argument structure; potential bases for *-able* must have both an external and a direct internal argument (*washable*, **snorable*). So among the productive affixes we must be able to distinguish different degrees of productivity.

Baayen (1989) develops a number of statistical measures for distinguishing productive from unproductive affixes and for gauging the degree of productivity of productive affixes. It is to these measures that we now turn.

2. Measuring morphological productivity

The relevant facts in the literature amount to the observations that a lesser degree of productivity is correlated with a higher mean token frequency, and to the fact that the frequency distributions of simplex (underived) types are less skewed, and contain fewer rare types, than the frequency distributions of productively coined formations. A measure of productivity should do justice to these obervations and should also meet the requirements that

1. it reflect the linguist's intuitions concerning productivity,
2. it express 'the statistically determinable readiness with which an element enters into new combinations' (Bolinger 1948: 18), and
3. it take into account that semantically or formally idiosyncratic words have the effect of lowering the value of the productivity measure.

A measure which satisfies these requirements is[8]

(2) $$\mathscr{P} = \frac{n_1}{N},$$

where n_1 is the number of types with the relevant affix occurring exactly once in the sample (the so-called hapax legomena) and N the total number of tokens of all words with that given affix. Broadly speaking, $\mathscr{P}$ expresses the rate at which new types are to be expected to appear when N tokens have been sampled. In other words, $\mathscr{P}$ estimates the probability of coming across new, unobserved types, given that the size of the sample of relevant observed types equals N. (Note that there are two kinds of samples involved here, the corpus itself, which is hoped to be a representative sample of the way language is put to use, and the individual samples of types with a particular morphological constituency that are found in the corpus. The figure N of [2] denotes the size of the individual samples.)

In order to understand what property of the word-frequency distributions $\mathscr{P}$ is exploiting, consider Figure 1. On the horizontal axis the frequency rank r is displayed. On the vertical axis, one finds the fraction of the n_r types with frequency r on the total number of types V. The n_r values themselves have been added for each bar. Only the first 15 ranks are shown. Thus we find that the 77 hapaxes in *-ness* account for 15.5 per cent of all different types V, that the 56 types that occur twice represent 11.2 per cent of V, etc.

Characteristic for productive WFRs is the way the word-frequency distribution is highly skewed to the left. In the case of *-ness*, the mode, that value of r for which n_r is greatest, is at the left-hand edge. Moreover, the types that occur once only in the sample represent a sizeable portion

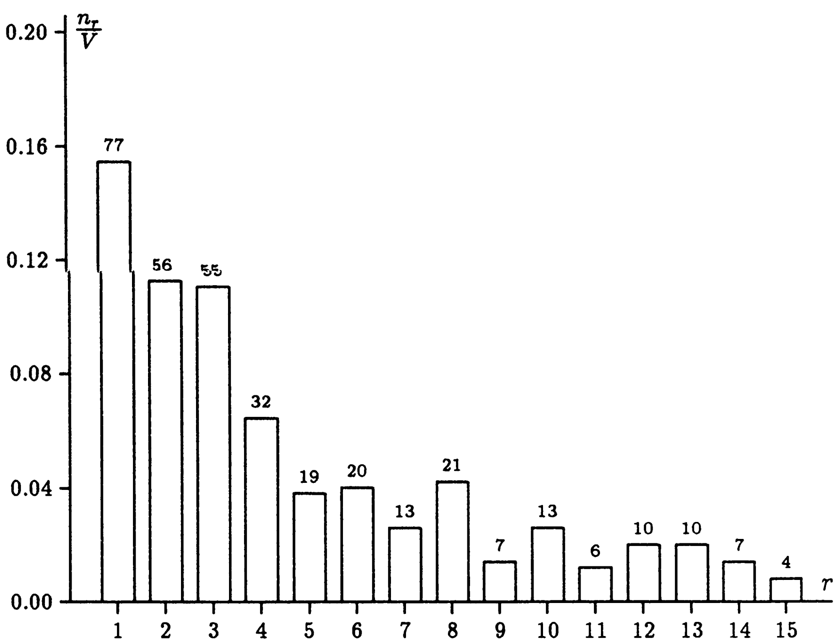

Figure 1. *The head of the frequency distribution of* -ness *(horizontally the frequency of the types* r *is displayed; vertically the fraction of types with the frequency* r *is shown; the absolute numbers of types* n_r *with the frequency* r *are added above the histogram bars)*

of all types. This means that roughly one in eight types in *-ness* occurs only once in the 18 million corpus used here. We may contrast this with the class of simplex nouns, where the hapaxes represent only 0.039 per cent of the types. Also, for some unproductive classes it is found that the mode occurs not at 1 but at some higher-valued rank. It is this difference in the shape of the word-frequency distributions that is exploited by $\mathscr{P}$.

The main interest of $\mathscr{P}$ is not that it is a descriptive statistic (like, for example, the mean token frequency, which is, of course, applicable to such frequency distributions as displayed in Figure 1) that summarizes the frequency data in such a way that frequency distributions are ranked in what we intuitively feel to be the right order of productivity. Rather, $\mathscr{P}$ expresses, be it in the language of mathematics, in a very real sense the linguistic notion of productivity, which broadens its scope of usefulness from a descriptive statistic to an analytic tool. To see this, consider the kind of information that is provided by $\mathscr{P}$. By making use of the information present in the type and token frequencies in a sample, $\mathscr{P}$ predicts at what rate new types, types that are not represented in that sample, will appear when we decide to enlarge the sample. Building on what is actually present in the sample, $\mathscr{P}$ makes a statement of what potentially could have been in the sample but has not been actualized in the sample for some reason or other. If the sample on the basis of which $\mathscr{P}$ is calculated faithfully reflects the properties of the population it is supposed to represent, $\mathscr{P}$ can be viewed as a measure of the potentiality of the word-formation process which underlies the sample. In this sense, $\mathscr{P}$ is a mathematical formalization of the linguistic notion of morphological productivity. We will not elaborate on the statistical derivation of $\mathscr{P}$ here, but some comments on what $\mathscr{P}$ is cannot be avoided, if this statistic is to be understood and used correctly.

When a corpus of words is compiled, we may consider this process of compilation as a sampling process in which new word types appear successively, and in which some types will be sampled more frequently than others. As we continue to increase our sample, the total number of tokens sampled, N, will increase. (Note that N is in fact the sample size.) Similarly, the number of word types sampled, to which we shall refer as V, will increase, but not at the same rate as N. In fact, we can plot the number of different types V obtained at the various stages of compilation against the size of the sample N at that stage of compilation, as illustrated by the curve in Figure 2. In other words, the number of types V can be considered as a function of the number of tokens N. For small values of N, $V(N)$ will first increase rapidly, but as more and more types have appeared, the rate at which new types occur will decrease. It is this rate

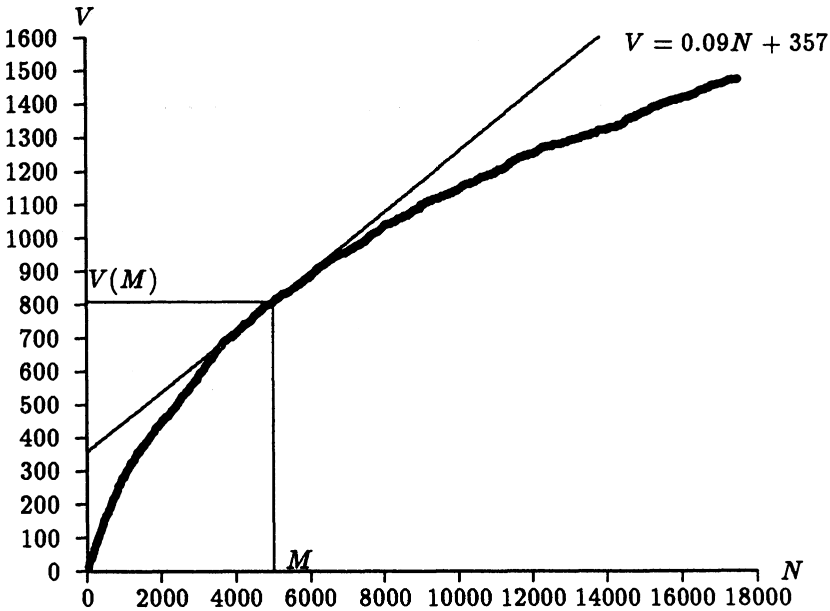

Figure 2. $\mathcal{P}$(M) *as the slope of the tangent to the growth curve of* V *(the curve shown here is that of the third singular present tense flectional* -s *as observed for a random sample of 17,481 tokens from the Cobuild corpus; the slope of the tangent in the point* (M = *5000,* V(M) = *807), 0.09, equals* $\frac{n_1}{N} = \mathcal{P}$ *for* N = M)

at which new types occur that is expressed by $\mathcal{P}$. In fact, given that the size of our sample is M, the value of $\mathcal{P}$ calculated for that sample, $\mathcal{P}(M)$, is the slope of the tangent to the growth curve of V in the point (M, $V(M)$). If $\mathcal{P}(M)$ is large, the growth curve of V will be very steep for $N = M$, indicating that many types remain to be sampled (the affix sampled is productive). In contrast, if $\mathcal{P}(M)$ is very small, the growth curve of V will be flattened out, which tells us that few if any new types remain to be sampled (the affix sampled is unproductive). Two points should be noted: first, that, like V, $\mathcal{P}$ is a function of the sample size N, and second, that V and $\mathcal{P}$ are functions that are related to each other (as primitive and derivative).

At this point we pause to consider under what conditions a type can be said to be 'new'. Obviously, a formation that has been recently created and that has not found its way into the established vocabulary of the speech community is a 'new' type. We will refer to such new types as neologisms. In the present corpus-based study of types, the notion of 'newness' is also used in two slightly different ways. First, when one is going through a corpus from beginning to end, the 'new' types are those that have not been encountered before. Note that what is a new type at sampling stage t_1 need not be a new type any more at stage t_2. When the end of the corpus has been reached, the hapaxes at this final stage t_f represent its 'new' types. Of course, it will depend on the size of the corpus whether these hapaxes will also be new in the sense that they are new to the speech community, that is, that they will be neologisms. A small corpus will tend to sample only words that are frequently used in the language, hence its 'new' types, or better, its hapaxes, will be 'old' items with respect to the language community. However, large corpora may contain types that are innovative with respect to the speech community's established formations. For instance, the 40,000,000 Dutch corpus made accessible by Celex contains tens of very low frequency types in the suffix *-heid* (the equivalent of English *-ness*) that are not registered in the most comprehensive standard dictionary of Dutch, van Dale (1976). Although some of these items are new only in the lexicographic sense in that they look familiar enough, others are true neologisms.

Second, the notion 'new' bears on the case when we use $\mathscr{P}$ to predict the rate at which 'new' types, new in the sense that they have not been observed in the corpus on the basis of which $\mathscr{P}$ is calculated, are expected to appear. If the corpus is large enough, and if it faithfully reflects the way in which words are put to use in the language community, a significant proportion of these 'new' types may again well be real neologisms. It should be stressed that when we interpret $\mathscr{P}$ as a measure of the rate at which new types are expected to appear, we use the word 'new' in this last sense. That is, we do not claim that all n_1 hapaxes counted in the sample are neologisms. Some may be neologisms, but this is not the point we want to make. The crucial idea is that $\mathscr{P}$ sheds some light on the extent to which the types that appear in the sample exhaust the available number of potential types in the population. When $\mathscr{P}$ is large, many types remain to be sampled. When $\mathscr{P}$ is small, nearly all types have been sampled at least once. In the former case, we are dealing with a productive process, for which a large, perhaps infinite number of possible types is characteristic. In the latter case, we are dealing with an unproductive process, where the number of types is small and, of course, finite.

Another question that is relevant here is whether it is possible for neologisms to be hapaxes. It might be argued that neologisms are typically coined to fulfil some need, and that this need will not be filled by the creation of a new lexeme if that lexeme is going to be used only once. In other words, might it not be the case that neologisms occur in clusters rather than once only, while the words that do occur only once are the rare words of the language? This would imply that the hapaxes cannot be the 'new' types of interest, which would invalidate our theory. Fortunately, this line of reasoning can be shown to be wrong. First, although many new words appear in clusters in the sample, thus giving rise to at least some of the types that occur twice or three times, etc., only, it is a simple fact of life that substantial numbers of regular, morphologically complex neologisms remain that are used only once. Why might this be the case? Interestingly, Kastovsky (1986) has pointed out that word formation may serve two different functions, not only what he calls labeling, but also syntactic recategorization. Labeling serves to designate segments of extralinguistic reality. In fact, it is far from clear that labeling has to result in the clustered use of neologisms — see, for example, Downing's (1977) deictic compounds of the type *apple juice chair*, or the following example from Kastovsky (1986: 594):

The Time Patrol also had to *unmurder* Capistano's great-grandmother, *unmarry* him from the pasha's daughter in 1600, and *uncreate* those three kids he had fathered.

Although it is possible that words for new concepts will be used more than once in a text, this is more likely to occur for typically referring expressions than for adjectives or verbs. Even more important, however, is Kastovsky's observation that word formation is also used for syntactic recategorization, with the aim of condensing information, introducing stylistic variation, and supporting text cohesion. Consider one of his examples (1986: 599):

If I were to attempt shadowing anybody, the *shadowee* would find himself as inconspicuous as though he were to walk down Piccadilly pursued by the Albert Memorial.

where *shadowee* is itself a partial repetition of *shadowing* and is not coined as a new concept that is going to be the topic for the next paragraph or so. Given this use of neologisms, it is not at all self-evident that neologisms cannot appear singly in texts. A second weak point of the above line of reasoning concerns the fact that it is a priori unclear why, if someone invests the effort to dredge up a so-called rare word from memory, this rare word should not be subject to the same clustering phenomenon as productively coined neologisms.

Having outlined the basic ideas underlying $\mathscr{P}$, we now turn to its linguistic interpretation. To our mind, the linguistic interpretation of $\mathscr{P}$ as the growth rate of V is that it expresses the degree of productivity of a word-formation process. A large number of hapaxes positively influences the value of $\mathscr{P}$. Similarly, the absence of large numbers of very high-frequency words also contributes to a high value of $\mathscr{P}$, since high-frequency words contribute many tokens to the sum of all tokens N. Hence a distribution with few high-frequency types is more likely to have a higher value of $\mathscr{P}$ than a distribution with many high-frequency types. These two requirements for a high $\mathscr{P}$ and hence for a high degree of productivity are met precisely by the more productive word-formation processes. These processes show up with the frequency distributions with the greatest degree of skewing in favor of low-frequency types. Conversely, the classes of simplex formations are characterized by large numbers of hard-worked, high-frequency words and only small numbers of hapaxes. These classes, which are skewed to a far lesser degree, show up with extremely low values of $\mathscr{P}$, a natural result, given that simplex classes are on the borderline of productivity by definition. (Although new simplex items are coined, they are formed on the basis of linguistic creativity rather than on the basis of word-formation rules. Moreover, since the different kinds of objects [rather than the properties of these objects] we encounter in daily life are Zipf-like distributed and show up with frequency distributions similar in shape to the word-frequency distribution displayed in Figure 1, whether we are dealing with biological species in a particular habitat or with anorganic substances dissolved in seawater, the frequency distributions of the words we use to denote these objects will owe their shape at least in part to the properties of the natural world. Hence the simplex items present a base-line condition for the assessment of productivity: a frequency-based analysis of morphological productivity should be sensitive to what the morphology adds to a distribution rather than to what is already intrinsically there given processes of creative coining and the distributional properties of the objects our words refer to.) Unproductive processes, finally, have frequency distributions which are highly similar to those of the simplex classes and, not surprisingly, are characterized by very low degrees of productivity.

At this point we may pause to consider the advantage of the statistic $\mathscr{P}$ above a summary statistic such as, for example, the arithmetic mean used by Anshen and Aronoff (1988). Without denying that mean and variance summarize important properties of word-frequency distributions, they do not by themselves disclose anything about the type richness of the population sampled. Differences in the mean token frequencies of *-ness* and *-ity*, as discovered by Anshen and Aronoff (1988), suggest that high token frequencies may be characteristic of the less-productive word-formation processes. Why this should be so cannot be clarified on the basis of what the arithmetic mean is — in fact the mean is simply a short way of stating the observation that there is a difference, even though the difference may suggest a link with productivity to the observer. In contrast, $\mathcal{P}$ can be used not simply to discover some frequency-related difference between samples, but to discover exactly those differences that relate to differences in the type richness of the populations involved. As an expression of the rate at which new types occur, $\mathcal{P}$ is the frequential reflex of the degree of productivity, and as such both observationally and analytically superior to the sample mean. While the argument that token frequencies are irrelevant to productively formed items in, for example, *-ness* (Anshen and Aronoff 1988) may have some initial appeal when sample means are used, a more realistic analysis becomes feasible in terms of $\mathcal{P}$. The preponderance of low-frequency types in the frequency distribution of *-ness*, which causes $\mathcal{P}$ to assume a high value, is itself indicative of a high degree of productivity since, especially in the case of the hapaxes in large corpora, the likelihood that we are dealing with formations that are NOT listed in the mental lexicon, and for which the availability of a word-formation process is crucial to their use, increases (see Baayen forthcoming). But now we are not forced to advance the unrealistic argument that, for example, the high-frequency types in *-ness* listed in Table 2 are not stored, in order to explain the differences in the mean frequencies observed for *-ness* and *-ity*.

It is important to observe that, as a measure of the degree of productivity, $\mathcal{P}$ is a relative measure that, by itself, cannot be used to make the categorial decision whether a rule is productive or not. When it is necessary to decide, not whether some affix is more productive than some other affix, but whether it is productive at all, we can make use of the fact that simplex words are on the bottom line of productivity and use the $\mathcal{P}$ value of the relevant set of simplex words to weight the affixal value of $\mathcal{P}$. For instance, we may compare the degree of productivity of some suffix, say *-ness*, with the $\mathcal{P}$ value of the corresponding class of simplex words, in this case the class of simplex nouns. Given the variance of n_1 for both *-ness* and the class of simplex nouns, we can test whether the 'new' formations in *-ness* have a probability of occurring that is significantly larger than that of 'new' simplex words. It is only for the productive classes that a positive difference is expected. In other words, if it would be easier to coin an entirely new simplex noun rather than to form a new noun in *-ness*, we could take *-ness* to be unproductive. Since in the case of *-ness* the reverse obtains, our method confirms that *-ness* is productive.

Of the three conditions for a useful measure of productivity, the condition that such a measure express the statistical probability with which new types occur is satisfied. In section 3 we show that the condition of a correlation with intuitions is also satisfied. With respect to condition three, we may note that, since semantically or formally idiosyncratic words typically turn out to be high-frequency items, they negatively affect the degree of productivity of the word-formation rule. Large numbers of such idiosyncratic formations will cause the degree of productivity to tend to zero.

Up to now, we have focused on productivity in the strict sense, namely, as the aspect of potentiality of word-formation rules. Of course, the notion of productivity can also be understood in a less-specific way when the numbers of different types are the main object of interest. Although V is, by itself, not a measure of potentiality or degree of productivity, it is an indicator of the extent of use and as such of interest to, for instance, the lexicographer. Moreover, since $V^{(N)}$ and $\mathscr{P}^{(N)}$ are related to each other as primitive and derivative, as shown in Figure 2, the status of V is in need of clarification. We therefore first comment on the relation between $\mathscr{P}$ and V and then suggest a tentative linguistic interpretation of V.

Although V and $\mathscr{P}$ are intimately linked, the fact that $\mathscr{P}$ is a function of N has as its consequence that the growth curve of a given affix is not fully characterized by $\mathscr{P}$. For instance, for some fixed N, V, and $\mathscr{P}$ we do not know with what rate the growth curve of V will flatten out for larger samples. Hence, on the basis of $\mathscr{P}$ by itself nothing can be said about the absolute number of types which might be expected to surface in larger samples. Similarly, $\mathscr{P}$ by itself cannot be used to estimate S, the number of possible types. Nevertheless, S, which is in fact the limit of the growth curve of V for $N \rightarrow \infty$, is of interest both to the study of linguistic productivity (see for example, Aronoff's index of productivity [1]) and to lexicography, although for opposite reasons.

In lexicography and applied linguistics, for instance, second-language teaching, the high-frequency types are of primary interest. This has the effect that the majority of, if not all, unproductive items are listed in bilingual dictionaries, while only the more frequent productive formations will generally be accorded entries. Consequently, dictionaries can afford to be exhaustive for the finite numbers of types belonging to unproductive classes, whereas exhaustive listing is unattainable and far too costly for productive classes with large or perhaps infinite S.

On the other hand, within the context of the present study the fact that one productive WFR may give rise to substantially more types than another productive rule raises the question why this should be so and in what way this might be connected with its productivity. We suggest that the number of observed types V is determined by at least three, probably interacting, factors, namely (1) the pragmatic usefulness of the affix, (2) the semantic flexibility of the word-formation process, and (3) the number of base words satisfying the conditions on the word-formation rule. The term 'pragmatic usefulness' captures the notion that some word-formation processes have a wider range of uses than others. For instance, in Dutch the suffix *-erd* is used to coin slightly pejorative personal names from adjectives, such as *natterd*, 'a wet person', from *nat*, 'wet'. The use of *-erd* is, because of its meaning, severely restricted, mainly to informal oral contexts. Hence, even though it is judged to be productive by Schultink (1962), it shows up with only 31 types in the 40,000,000 Celex database of Dutch word forms based on the INL corpus (version N2.6). With respect to the semantic flexibility of word-formation rules, we suggest that for instance the fact that compounding is extremely productive both in terms of $\mathscr{P}$ and in terms of V is at least in part due to the semantic versatility of this word-formation process.[9] Finally, the effect of restrictions on word-formation rules is especially apparent when the frequency distributions of rival affixes, or the allomorphic variants of a single affix, are studied. In one sense the restrictions on a word-formation rule define the domain where a rule can be productive. From a slightly different point of view, such restrictions, when they strongly limit the number of available base words, may, in combination with a high pragmatic usefulness or semantic versatility, have the effect of concentrating the use of a word-formation rule to a relatively small number of types, thereby lowering the degree of productivity. Whatever the precise interaction of these factors may be, they all play a part in determining the number of types V, which we suggest is a measure of the extent of use of a WFR, and they all enter into the assessment of morphological productivity in some more general sense. We will refer to this more general sense of productivity as global productivity. The global productivity P^* of a WFR can be summarized in terms of its coordinates in the $\mathscr{P}$-V plane, with the degree of productivity on the horizontal axis and the extent of use V on the vertical axis, as shown in Figure 3. The globally more productive rule will have large values for both V and $\mathscr{P}$, the globally unproductive rule will show up with few types and a low-valued $\mathscr{P}$.

This two-dimensional analysis of P^* of Figure 3 has, unfortunately, the drawback that it remains difficult to assess which WFR is the globally more productive one when differences along both dimensions are involved. For instance, *-ness* is clearly more globally productive than deadjectival *-ian*: although both affixes are characterized by approximately the same degree of productivity, *-ness* has a far larger V. For

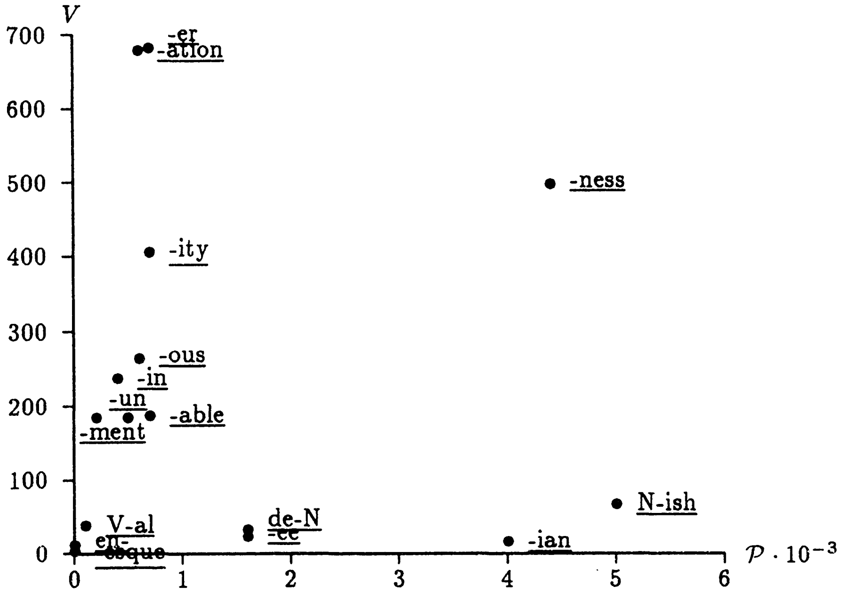

Figure 3. *Global productivity* P* *for various affixes (the degree of productivity* 𝒫 *is found on the horizontal ordinate; the vertical ordinate represents the extent of use* V)

larger corpora, the absolute increase in types is predicted to be greater for *-ness* than for *-ian*. (Consequently, an exhaustive listing of known types in *-ian* in the dictionary is feasible and realistic, even though this suffix is productive.) But what about *-er* and *-ness*? It is simply impossible to gauge on the basis of V and 𝒫 which WFR is the more productive one: V and 𝒫 do not contain enough information for predicting how the shape of the curve of V will develop when N is increased.

For a more precise evaluation of P^* other methods are available, however. Unlike 𝒫, which is a so-called nonparametric statistic that makes minimal assumptions concerning the properties of the underlying population, these so-called parametric methods make use of more elaborate theoretical models. Unfortunately, the more reliable models of this kind are far less easy to apply, but in principle they can be used to obtain an estimate of the theoretically possible number of types S, and once S is known, the actual vocabulary V can be compared with the potential vocabulary S in order to evaluate P^*. Section 4 sketches how an analysis of P^* along these lines might proceed. The main aim of the present paper is, however, to show that one can already come a long way on the basis of 𝒫 and V.

Before turning to the quantitative analysis of the productivity of a number of English affixes, it should be noted that the method proposed here is constrained by the corpus it is applied to. The corpus establishes a frame of reference for the comparison of the productivity of the affixes it contains. Hence, if comparisons are to be made across corpora, these corpora should be compatible with respect to both the kind of texts sampled and their size.[10]

3. Productivity of selected affixes

3.1. *General remarks*

The affixes we have chosen to investigate are the following. For noun-forming affixes we have chosen the agentive/instrumental suffix *-er* (*baker, sweeper*) and the patient noun suffix *-ee* (*employee*); the process/result noun–forming suffixes *-ation* (*representation*),[11] *-al* (*refusal*), and *-ment* (*commitment*); the abstract noun–forming suffixes *-ness* (*happiness*) and *-ity* (*purity*); and the suffixes *-ian* (*comedian, civilian*) and *-ism* (*Marxism, purism*). Affixes which form adjectives are the suffixes *-ish* (*clownish, reddish*), *-ous* (*monstrous*), *-able* (*washable*), *-ive* (*impressive*), and *-esque* (*picturesque*); and the prefixes *un-* (*unsure*) and *in-* (*impure*).[12] Among the verb-forming affixes are the prefixes *de-* (*debug*), *en-* (*enchain, enlarge*), *be-* (*bespeak, befriend, belittle*) and *re-* (*rewash*); and the suffixes *-ize* (*finalize, hybridize*) and *-ify* (*codify, purify*). Tables 3–5 show the values for N (number of tokens), V (number of types), $\mathscr{P}$ (productivity), n_1 (number of types occurring once), and n_2 (number of types occurring twice) for the affixes listed above, as well as for the classes of simplex nouns, adjectives, and verbs. Within each subtable the affixes have been sorted according to the category of item they form and the category of item they attach to. Where a particular affix, for example *-ish* or *-ize*, can attach to bases of two different categories, it will appear twice in the appropriate table; in other words, we have calculated $\mathscr{P}$ separately for *-ish* which attaches to nouns and *-ish* which attaches to adjectives, and so on. Within each table, comparable affixes (that is, affixes which attach to and which form the same category of words) have been listed in order of decreasing $\mathscr{P}$.

There are three general observations we can make about the data in Tables 3–5. Note first that, as was mentioned in the preceding section, affixes may be represented by a relatively large number of types (V) and yet be ranked fairly low in productively ($\mathscr{P}$), and vice versa. Good examples of this are *-ee* versus *-al* and *-ment* in Table 3, or *-ish* and *-ous*

Table 3. *Noun-forming affixes*

Affix 18,000,000	N	V	$\mathscr{P}$	n_1	n_2
From verbs					
-ee	1213	23	0.0016	2	2
-er	57683	682	0.0007	40	40
-ation	74466	678	0.0006	47	37
-ment	44419	184	0.0002	9	7
-al	7317	38	0.0001	1	3
From adjectives					
-ness	17481	497	0.0044	77	54
-ian	505	16	0.0040	2	0
-ity	42252	405	0.0007	29	21
-ism	3755	82	0.0005	2	4
From nouns					
-ian	2898	27	0.0007	2	0
-ism	3290	50	0.0006	2	1
-al	29445	45	0.0001	2	0
Simplex nouns	2781258	6582	0.0001	256	257

Table 4. *Adjective-forming affixes*

Affix 18,000,000	N	V	$\mathscr{P}$	n_1	n_2
From verbs					
-able	15004	187	0.0007	10	8
-ive	21337	179	0.0003	6	8
From adjectives					
-ish	290	16	0.0034	1	2
un-	11952	184	0.0005	6	9
in-	14426	237	0.0004	6	6
From nouns					
-ish	1602	67	0.0050	8	4
-ous	21861	264	0.0006	13	10
-esque	238	3	0.0000	0	0
Simplex adjectives	994716	1659	0.0001	60	32

in Table 4. Affixes may be quite productive even if they show up relatively infrequently in a corpus. The crucial factor in productivity in the strict sense, that is $\mathscr{P}$, is that they show a relatively high proportion of hapaxes, and not that they have a large number of types.

Second, observe that rival affixes, that is, affixes which attach to and form words of the same category, and which have more or less the same semantic effect, rank according to $\mathscr{P}$ in the way our intuitions would lead

Table 5. *Verb-forming affixes*

Affix 18,000,000	N	V	$\mathscr{P}$	n_1	n_2
From verbs					
re-	23591	96	0.0000	1	3
be-	1662	19	0.0000	0	0
From adjectives					
-ize	14083	61	0.0001	1	0
-ify	7764	17	0.0000	0	0
en-	6705	11	0.0000	0	0
be-	82	1	0.0000	0	0
From nouns					
de-	1887	32	0.0016	3	1
-ize	12491	85	0.0002	2	2
-ify	9815	33	0.0000	0	1
en-	20961	40	0.0000	0	0
be-	706	5	0.0000	0	0
Simplex verbs	3660693	2581	0.0000	24	24

us to expect. In Tables 3–5 *-ness* ranks higher in $\mathscr{P}$ than *-ity*, *-ish* higher than *-ous*, *un-* higher than *in-*, and *-ation* higher than *-al* or *-ment*. Of course, there is more to be said about such pairs of rival affixes and about how the effect of restrictions on word-formation rules can be measured, and we will discuss a number of these cases in some depth in sections 3.2–3.7.

Third, a comparison of the $\mathscr{P}$ values of a derived class with that of the corresponding class of simplex words sheds light on the question whether the probability of having to process new types is greater for the derived class than for the simplex class. For instance, in the case of the suffix *-ness*, the difference in the values of $\mathscr{P}$, 0.0044 for *-ness* but only 0.0001 for simplex nouns, argues strongly in favor of the productivity of *-ness*. In contrast, *-al*, as in *refusal*, is clearly unproductive: its $\mathscr{P}$ value is identical to that of the simplex nouns. For the suffix *-esque* (*picturesque*) the situation is even worse, since there are no low-frequency types among the three adjectives in the 18,000,000 corpus. Our data strongly suggest that, contrary to Bauer's (1983: 224) claim that *-esque* 'is still productive', it is unlikely that speakers of English can form novel adjectives in *-esque* spontaneously and unintentionally.[13]

As an illustration of $\mathscr{P}$ as a measure of the degree of productivity of a word-formation rule that quantifies the potentiality of WFRs, consider privative *de-* (*delouse*, productive) and *en-* in the case where it attaches to nominal base words (*encourage*, unproductive). As shown in Table 6, productive *de-* shows up with $\mathscr{P} = 0.261$ in the 1,000,000 Kučera–Francis

Table 6. *Comparison of 𝒫 and* V *for the Kučera–Francis and Celex (Cobuild) corpora:* de- *and* en-

	Kučera–Francis	Celex
de-		
N	23	1887
V	13	32
𝒫	0.261	0.0016
en-		
N	421	20961
V	36	40
𝒫	0.021	0.0000

corpus and retains a value of 0.002 in the 18,000,000 Celex database. The number of types in the larger corpus is roughly 2.5 times that of the smaller corpus. In the case of *en-* a different picture emerges. In the Kučera–Francis corpus we find 36 types in combination with an already low value of 𝒫, 0.021, compared to that of *de-*, 0.261. In the larger Celex database the number of types is only slightly raised from 36 to 40, while the 𝒫 value drops to zero. Although the present comparison should be treated with caution — differences between British and American English, as well as differences in the materials sampled for the two corpora may blur the picture presented here — it illustrates two important properties of 𝒫, namely that it is a good qualitative predictor of the number of new types which may be expected for larger samples, and that the value of 𝒫 decreases for increasing sample size.

In order to allow a comparison to be made with a morphological process that is generally considered to be fully productive, we finally discuss two inflectional endings, the third-person singular present-tense marker *-s* and the plural marker *-s*. The former shows up in 674,183 tokens, a total of 4,094 types, 609 of which are hapaxes, yielding a 𝒫 value of 0.0009. The latter is observed for 654,893 tokens that can be traced to 9,728 types, among which 1,280 hapaxes, so that 𝒫 is 0.0020. That such high values of 𝒫 are found for such extremely large samples with over 650,000 tokens is a clear indication of their enormous productivity. (To see this, consider the hypothetical case in which we stop sampling the third-person singular present-tense marker *-s* when 17,481 tokens have been obtained, the value of *N* for derivational *-ness*. For this much smaller sample, *V* equals 1,474 and 𝒫 0.0348, which is very much larger than the 𝒫 of 0.0044 observed for *-ness*, as expected.)

3.2. *The rival suffixes* -ness *and* -ity

-ness and *-ity* are both suffixes which form abstract nouns from adjectives. According to Marchand (1960: 271), although *-ness* 'may be tacked on to any adj, those of native stock form the majority'. *-ity*, however, only rarely goes on native bases, its use being more or less confined to the latinate segment of the vocabulary. Anshen and Aronoff (1988: 645) show that it is much easier to create nonce forms in *-ness* than it is to create nonce forms in *-ity*; when asked to list all forms in *-ibleness*, *-ibility*, *-iveness*, and *-ivity* that they could think of in 90 seconds, their subjects[14] created a total of 12 and 16 nonce forms respectively for *-ibleness* and *-iveness*, 8 and 9 respectively for *-ibility* and *-ivity*. All of this suggests informally that *-ness* is a more productive affix than *-ity*.

This is generally what our measures show, although the picture is worth looking at in some detail. Since *-ness* has a $\mathscr{P}$ value of 0.0044, and *-ity* shows a value of 0.0007, *-ness* is obviously the generally more productive affix. This is not to say that *-ity* is unproductive, however. First, when compared with the set of simplex nouns, both *-ness* and *-ity* show up with values of $\mathscr{P}$ which are far higher than that of simplex nouns: 0.0044 and 0.0007 against 0.0001.[15]

Second, with respect to certain sorts of bases *-ity* is clearly dominant. Again, we need not limit ourselves to a simple count of types (V) when studying the effect of the kind of base word on the productivity of the word-formation rule. Instead, we can in addition calculate the $\mathscr{P}$ value for each morphological subdomain of the rule. The histograms in Figures 4 and 5 summarize the frequency distributions for *-ness* and *-ity*, where we have partitioned the set of formations in these suffixes into a number of subsets according to the properties of the base words involved. Using van Marle's (1985) terminology, we can say that these histograms summarize the frequency characteristics of the derivational subdomains of *-ness* and *-ity*. For each subdomain, the number of types occurring exactly once (n_1) or twice (n_2) is listed, as well as the numbers of types $n_{r \in I}$ in the frequency intervals (I) 3–10, 11–100, and 101 plus. We will use these histograms to gain some insight into the weight of the various restrictions in force for the suffixation of *-ness* and *-ity*.

With respect to *-ness*, we find that it shows the larger numbers of types at the low frequencies (types occurring once or twice) for simplex word forms. It alone attaches to complex forms in *-ed*, *-ful*, *-less*, *-some*, *-ish*, *-y*, and *-ly*, all native affixes. *-ity* shows up on words in *-ic*, *-al*, and *-able*/*-ible*; with respect to the latter, it shows greater strength than *-ness* in all frequencies, especially the low ones, which suggests that in this domain *-ity* is more productive than *-ness*. Both *-ness* and *-ity*, finally,

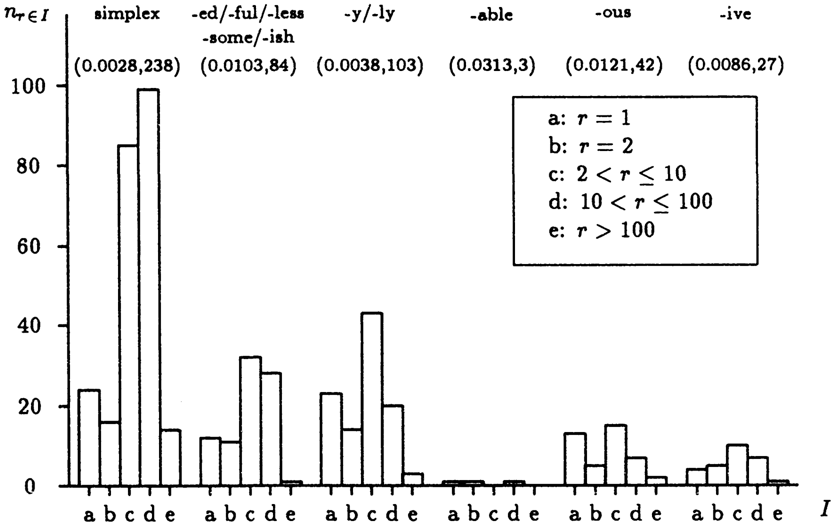

Figure 4. *Summary of the frequency distributions of the derivational subdomains of* -ness *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* $n_{r\in I}$ *for which the frequency rank* r *falls within* I; *for each subdomain* $\mathscr{P}$ *and* V *have been added)*

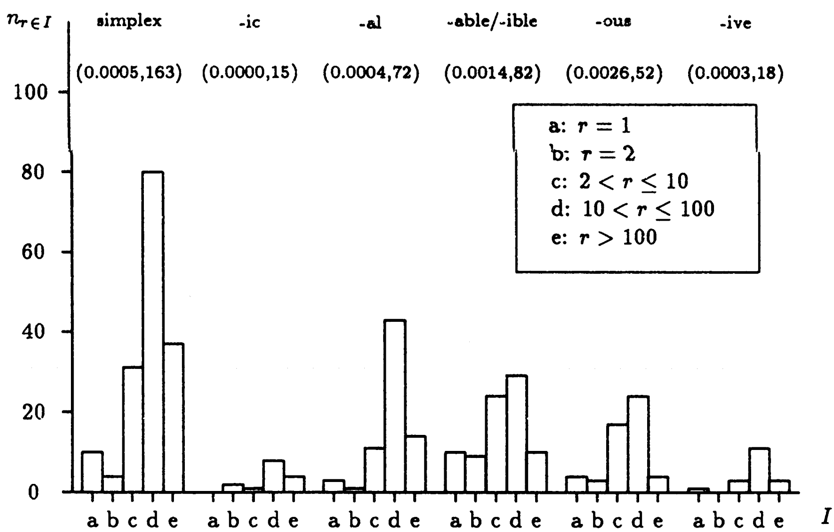

Figure 5. *Summary of the frequency distributions of the derivational subdomains of* -ity *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* $n_{r\in I}$ *for which the frequency rank* r *falls within* I; *for each subdomain* $\mathscr{P}$ *and* V *have been added)*

attach to words in *-ive* and *-ous*. With respect to *-ive*, *-ness* shows, apart from a higher number of types, more low-frequency occurrences with *-ive* than *-ity* does, suggesting that it is the more productive suffix in this domain. Turning to *-ous*, note that a simple count of types would suggest that *-ity* is the stronger affix in this domain. Since *-os-ity* is, because of the lowering of the vowel of *-ous*, less transparent than *-ous-ness*, we would here have a counterexample against Cutler's (1980) observation that the more transparent affixes are the more productive ones. Inspection of the $\mathscr{P}$ values shows, however, that the study of productivity in terms of type frequencies only may be misleading: although *-ous-ness* shows up with fewer types than *-ous-ity*, it has the higher degree of productivity.

Finally note that what the histograms in fact demonstrate is that there is a sort of paradigmatic aspect (in the sense of van Marle 1985) to the pattern of productivity of these rival affixes. To some extent the rivals divide up the range of possible bases and show productivity in disjoint segments of this range.

3.3. *The suffixes* -ish *versus* -ous

The suffixes *-ish* and *-ous*, like *-ness* and *-ity*, are rival affixes which divide up their range of bases along the familiar native-versus-latinate lines. For *-ish* attached to nouns, the Celex corpus lists such forms as *apish*, *doltish*, *foppish*, *wolfish*, *biggish*, and *foolish*, and for *-ous* such forms as *humorous*, *vaporous*, *carnivorous*, *leprous*, *idolatrous*, *adventurous*, and *tumultuous*. Our measures confirm the intuitive feeling that *-ish* is more productive than *-ous*; *-ish* shows $\mathscr{P}=0.0050$; *-ous* has $\mathscr{P}=0.0006$. For both affixes the values of $\mathscr{P}$ exceed that of the simplex adjectivs, 0.0001, with the size of the difference being in accordance with the greater productivity of *-ish*.

3.4. *The prefixes* un- *versus* in-

Our own intuitions would lead us to expect that the negative prefix *un-* should be more productive than its rival *in-*. Both Marchand (1969) and Zimmer (1964, quoting Jespersen) share these intuitions. Marchand claims that 'With adjs the stronger rival has been native *un-* which is ousting *in-* more and more' (1969: 120). Jesperson comments, 'It should be noted that while most of the *in-* words are settled once and for all, and have to be learned by children as wholes, there is always a possibility of forming new words on the spur of the moment with the prefix *un-* ...' (Zimmer 1964: 28).

Still, both Zimmer and Marchand are aware that there are restrictions on the productivity of *un*-. As Zimmer and others before him have pointed out, *un*- tends not to attach to any base which is semantically negative (for example, **unbad*, **unnaughty*,**unsick*). And *in*-, as Marchand comments, has some productivity in the sphere of learned and scientific (presumably latinate) bases. Zimmer (1964: 29) in fact quotes the form *immanageable* as this sort of new coinage in *in*-.

Again, our measures of productivity accord nicely with these intuitions and informal observations. *Un*- shows $\mathscr{P} = 0.0005$, for *in*- $\mathscr{P} = 0.0004$. Both values exceed that of the simplex adjectives. What these figures suggest is that although *un*- is more productive than *in*-, both the restrictions on *un*- and the ability of *in*- to attach to learned and scientific bases bring the $\mathscr{P}$ values of these two closer together. In fact what we find when we take a closer look at the Celex database is another case where rival affixes divide up the range of possible bases (here adjectives) into sets which overlap very little. The histograms in Figures 6 and 7 illustrate this nicely. *Un*- shows greater productivity with complex bases in *-ed*, *-ing*, and *-ful*, and with adjectives in *-y* (all native affixes). *In*- shows greater strength with forms in *-able*/*-ible*/*-uble*, *-ive*, and *-ous*.

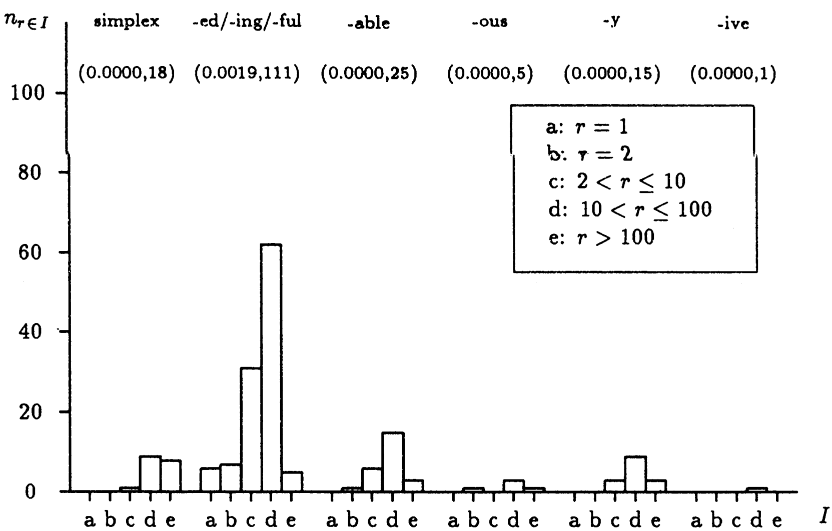

Figure 6. *Summary of the frequency distributions of the derivational subdomains of* un- *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* $n_{r \in I}$ *for which the frequency rank* r *falls within* I*; for each subdomain* $\mathscr{P}$ *and V have been added)*

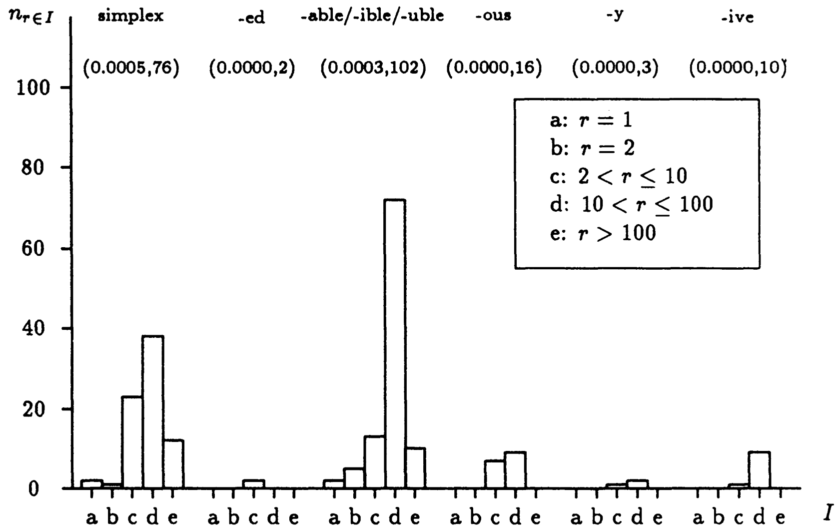

Figure 7. *Summary of the frequency distributions of the derivational subdomains of* in- *(on the horizontal axis the token frequency intervals I are displayed; on the vertical axis one finds the total number of types* $n_{r \in I}$ *for which the frequency rank* r *falls within* I*; for each subdomain* $\mathscr{P}$ *and* V *have been added)*

3.5. *Deverbal affixes 1:* -able, -er, *and* -ee

In this section we discuss the deverbal affixes *-able*, *-er*, and *-ee*, all of which we feel intuitively to be productive. Marchand's intuitions appear to accord with ours; he remarks that both *-able* and *-er* attach to verbal bases of all kinds, both native and latinate, and that *-ee* is a 'vogue' morpheme in American English (1969: 210), although some coinages in *-ee* '... have a playful nuance and a decidedly transitory character'. Of these affixes, however, *-able*, *-ee*, and *-er* are subject to restrictions having to do with predicate-argument structure (PAS) of the verbs to which they attach (see Levin and Rappaport 1986; Rappaport and Levin 1988 for discussion of PAS). Both *-able* and *-ee* attach only to verbs which have direct internal arguments (in the sense of Williams 1981). Thus, forms like *washable* and *employee* are possible, but not **snorable* or **snoree*. *-able* in addition seems to require that its bases also have external arguments, whereas *-ee* does not have this further requirement. For example, from an unaccusative verb like *arrive*, we can get *arrivee*, but not **arrivable*. *-er* appears also to have restrictions on the PAS of the verbs to which it attaches.[16]

The $\mathscr{P}$ values of these three affixes are as shown in Table 7. There are

Table 7. *Deverbal affixes: $\mathcal{P}$ and* V

18,000,000	V	$\mathcal{P}$
-ee	23	0.0016
-er	682	0.0007
-able	187	0.0007

a number of observations we can make about these findings. First, *-ee* seems to be the strongest in $\mathcal{P}$, which is at first sight somewhat surprising. However, for a 'vogue' morpheme a degree of productivity that is somewhat higher than expected makes sense, since the formations with a 'decidedly transitory character', which are, of course, low-frequency items, raise the degree of productivity. Also note that the high number of rare types, together with the low extent of use (23 types in an 18,000,000 corpus) may also be the quantitative reflection of the 'playful' character of these formations.

Second, observe that the $\mathcal{P}$ values for the affixes which attach to verbs are surprisingly low compared to noun- or adjective-forming affixes which attach to categories other than verb. For example, the $\mathcal{P}$ value of *-ish*, which attaches to nouns, is 0.0050, as opposed to 0.0007 for *-able*; and the $\mathcal{P}$ value of *-ness*, which attaches to adjectives, is 0.0044, as opposed to 0.0016 for *-ee*. There is a relatively simple explanation for the discrepancy in $\mathcal{P}$ values, however, that has to do with differences in the frequencies of use between nouns and adjectives versus verbs. Table 8 summarizes some relevant data on the type frequencies of nouns, verbs, and adjectives in the 18,000,000 database, together with $\mathcal{P}$ and the mean type frequency $\bar{f}=\frac{N}{V}$. What we observe is that verbs are used far more intensively than nouns or adjectives. The fact that verbs show up as the most intensively used items of the language has an important consequence for formations derived from verbs. Generally, the frequency of a derived item is less than that of its base word, that is, its derivation ratio is less than unity. Let's assume, for ease of exposition, that deverbal and denominal complex items have the same (mean) derivation ratio, say 0.5 derived

Table 8. *Nouns, verbs and adjectives in the Celex database*

18,000,000	V (simplex)	V (total)	$\mathcal{P}$	$\bar{f}$
verbs	2581	4964	0.000007	1418.32
nouns	6587	15199	0.000092	422.56
adjectives	1659	5428	0.000060	599.59

items per base word. Given that the frequency of a derived word will be half of that of its base word, we find that if this base word is a verb, the derived word's frequency will be roughly 700 in the light of Table 8. But if it is a noun, its frequency will be roughly 200. That is, other things being equal (which of course they are not), deverbal items may be expected to have higher frequencies of use than denominal items, and this causes $\mathcal{P}$ to assume lower values when the base word is a verb. We conclude that the large differences in $\mathcal{P}$ values between *-ness* and *-er*, *-ee* or between *-ish* and *-able* arise not so much from phonological or morphological restrictions on these rules, but from the categorial nature of the kind of base words selected.

3.6. *Deverbal affixes 2:* -ation, -ment, *and* -al

-ation, *-ment*, and *-al* are three of the suffixes which create process/result nominalizations from verbs and as such form another set of rival affixes. All have restrictions on the sorts of bases they attach to. According to Marchand (1969), *-ation* has the fewest restrictions; it attaches to simplex verbs, as well as to complex verbs in *-ify*, *-ize*, and *-ate*. *-al* attaches mostly to latinate bases, with the added restriction that the last syllable of the base must bear stress. *-ment* occurs often with verbs in *en-* and *be-* (*encouragement*, *bereavement*), but since neither of these patterns is productive (see section 3.7 below), we would expect its productivity to be quite restricted as well. The $\mathcal{P}$ values of these suffixes indicate that *-ation* is the most productive of the three ($\mathcal{P}=0.0006$), with *-ment* being barely productive ($\mathcal{P}=0.0002$) and *-al* on the borderline of productivity ($\mathcal{P}=0.0001$, exactly the $\mathcal{P}$ value for simplex nouns). This accords nicely with our intuitions. Note again that even the most productive of these three suffixes has a relatively low $\mathcal{P}$ value when compared with *-ness*, a noun-forming suffix which takes adjectives as bases. Again, we attribute this to the overall low frequency of verbs, as discussed in section 3.5.

3.7. *Verb-forming affixes*

The final group of affixes we will consider consists of the verb-forming affixes *de-*, *en-*, *be-*, *-ize*, *-ify*, and *re-*.[17] Of these affixes, our intuitions tell us that *de-*, *-ize*, and *re-* should be productive, *-ify*, *be-*, and *en-* less productive or even entirely unproductive. Marchand suggests that *-ify* affixes almost exclusively to non-native bases. *-ize*, according to Marchand, can attach to native bases and is especially productive in specialized areas of vocabulary. However, he notes that there are phonetic constraints on the attachment of *-ize*; *-ize* seems to favor bases ending in the segments [n,l,r,s]. It attaches less easily to bases ending in [t,d,m], occasionally to bases ending in [ǰ], and almost never to bases with other final segments. *De-* has no phonetic restrictions of this sort, and Marchand's judgment agrees with ours that *de-* is a live affix.

We find the $\mathscr{P}$ values listed in Table 9 for these six affixes. Note again that the $\mathscr{P}$ values are quite low, even for the more productive of these verb-forming affixes. (All $\mathscr{P}$ values are given to seven decimal digits in order to tease the truly unproductive and productive rules apart.) However, the $\mathscr{P}$ value of simplex verbs is far lower than even the low value for *re-*, while the truly unproductive affixes end up with $\mathscr{P}$ values lower than that of simplex verbs.

The values above accord roughly with our intuitions, although not entirely. *De-* appears to be the most productive, perhaps because it has fewer restrictions on attachment than *-ize*. *-ize* intuitively 'feels' productive, so we find its $\mathscr{P}$ value surprisingly low, although the phonological restriction which Marchand mentions might explain this low value at least partly. Also, since the Cobuild project was weighted toward general texts, rather than technical or scientific texts, the full productivity of *-ize* might not show up as well in Celex as it might. *-ify* and *en-* show up completely unproductive in the database. The histograms in Figures 8 and 9 clarify the picture even further; the bulk of items in *en-* and *-ify* are high-frequency items, whereas *-ize* and *de-* show a fair number of items in the low-frequency ranges.

Perhaps the only real surprise in the data is the $\mathscr{P}$ value for *re-*, which suggests, contrary to our intuitions, that *re-* is of a rather low degree of productivity. Although a comparison with the simplex verbs or with unproductive *be-* shows *re-* to be productive, *de-* reveals a difference in degree of productivity that does not accord with our intuitive feeling that *de-* and *re-* are roughly equally productive.

Table 9. *$\mathscr{P}$ values for verb-forming affixes to seven decimal places*

18,000,000	$\mathscr{P}$
de-	0.0015898
-ize	0.0000710
re-	0.0000423
-ify	0.0000000
be-	0.0000000
en-	0.0000000
simplex verbs	0.0000065

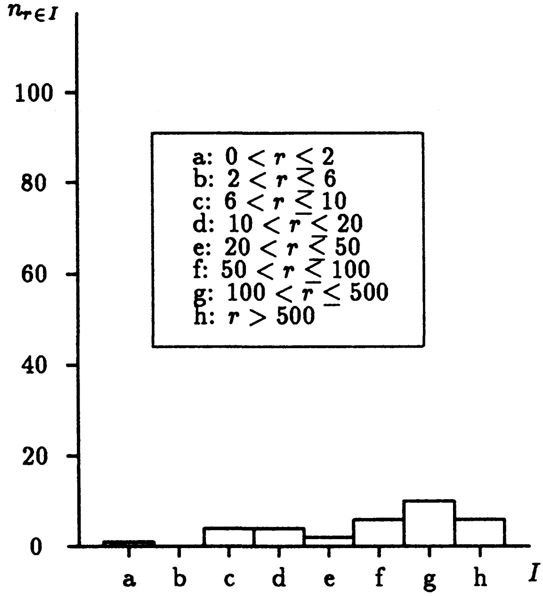

Figure 8. *Summary of the frequency distribution of* -ify *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* $n_{r\in I}$ *for which the frequency rank* r *falls within* I)

In order to gain some insight into the causes of the low value of $\mathscr{P}$ for *re-* we consider its frequency distribution in more detail (see Figure 10). The histogram of *re-* shows a number of low-frequency items in combination with a striking number of very high-frequency formations. A first observation then is that the presence of this high number of very high-frequency words is the main cause of the low value of $\mathscr{P} = \frac{n_1}{N}$. In other words, *re-* is atypical, not because of a lack of low-frequency items, but because of the presence of many high-frequency items, the default case for productive affixes being a relatively small number of such types, as shown in Figure 11 for *-ness*. We know of one other such case, namely comparative *-er* (*warmer*), which shows up with the rather low $\mathscr{P}$ value of 0.0016 ($V=354$, N=46698), in spite of being arguably an inflectional morpheme which we would expect to be productive. Interestingly, we find that the bulk of the tokens (approximately 30,000 of the 46,698 tokens) is accounted for by the ten most frequent items, with suppletive *better* as the formation with maximal frequency. In fact, the distribution shows quite a large number of hapaxes (77) and a sizeable number of types occuring twice (39). For our data on *re*, the ten most frequent types account for roughly 15,000 of the 24,000 tokens. On examination many

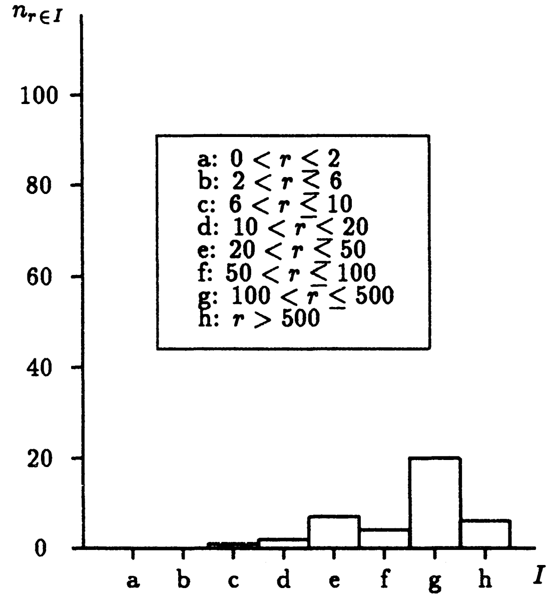

Figure 9. *Summary of the frequency distribution of* en- *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* n_{rI} *for which the frequency rank* r *falls within* I)

of these high-frequency words appear either to be noncompositional or to have the noncompositional meaning as the preferred reading. For instance, the verbs

remove	5474
recover	2085
recall	2295
reinforce	1107
react	1054
review	1007
recite	369

do not transparently encode the meaning 'repeat the action expressed by the verbal base', even though their meanings may, by various schemes of reasoning, be related to the meaning of the prefix *re-*. In the case of *remove*, for instance, the first reading that comes to mind is a separative one, 'move away', rather than the repetitive one, 'move again'. The typically compositional readings are found for the following low-

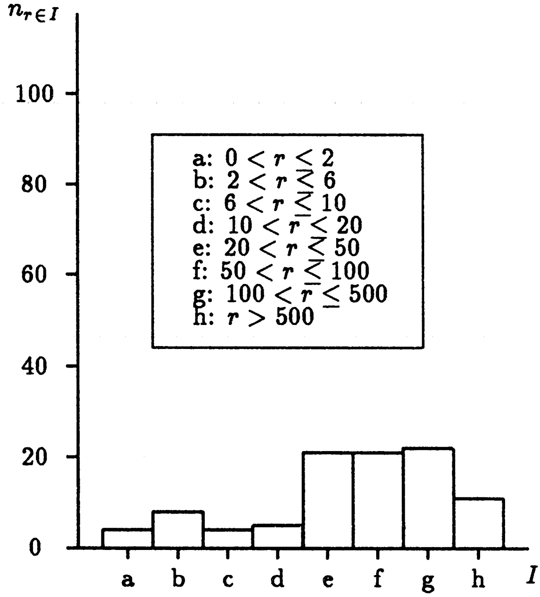

Figure 10. *Summary of the frequency distribution of* re- *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* $n_{r \in I}$ *for which the frequency rank* r *falls within* I)

frequency words:

reheat	1
reforest	2
repoint	2
retread	2
resole	3
retake	4

When we eliminate the high-frequency lexicalized formations from the frequency distribution of *re-*, we find that $N = 10200$, $V = 90$, $\mathscr{P} = 0.000098$, $n_1 = 1$, $n_2 = 3$. This illustrates the fact that $\mathscr{P}$ satisfies the third requirement discussed in section 3, namely, that taking into account semantically or formally idiosyncratic words has the effect of lowering the degree of productivity. With the new value of $\mathscr{P}$, 0.0001 to four decimal digits, the degree of productivity of this suffix is more in harmony with our intuitions that *re-* is a productive affix. Although the large difference in degree of productivity in comparison with *de-* is not resolved, we have shown that the atypical frequency distribution of *re-* is not incompatible with its productivity. In addition, we suspect that the

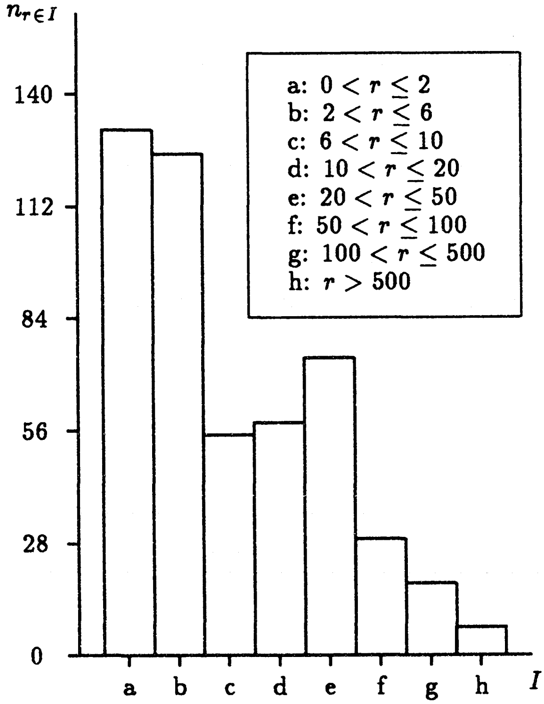

Figure 11. *Summary of the frequency distribution of* -ness *(on the horizontal axis the token frequency intervals* I *are displayed; on the vertical axis one finds the total number of types* $n_{r \in I}$ *for which the frequency rank* r *falls within* I)

different semantics of the two prefixes[18] renders a quantitative comparison in terms of $\mathscr{P}$ unfeasible — note that the more neutral comparison with simplex verbs has shown that both prefixes are unquestionably productive. As a final remark, we may point out that in terms of the overall productivity, P^*, the prefixes *re-* and *de-* are on a par, with *de-* evidencing the higher degree of productivity and *re-* the higher number of types (32 versus 90 respectively).

From a methodological point of view this study of the productivity of verb-forming affixes is of interest in that it stretches a $\mathscr{P}$-based analysis to its limits. For instance, the productive verb-forming affixes appear with hapaxes in the 18,000,000 corpus studied here, but their numbers are so very low that it is unlikely that we would still observe any hapaxes for verb-forming affixes in, say, a 40,000,000 corpus. The fact that for productive *re-* the number of types occurring twice is larger than the number of hapaxes ($n_2 = 3$, $n_1 = 1$) points in the same direction. But, even though $\mathscr{P}$ is expected not to be of much use for very large samples (see the next section for some discussion of why this should be so), we expect that productive and unproductive categories can still be distinguished in much the same way as *en-* and *re-*. Nevertheless, we are left with a final question, namely why derived verbs appear with so few hapaxes where comparable derived adjectives and nouns show up with healthy values for n_1.

There are two factors which may be relevant here. First, compared to nouns and adjectives, verbs are very hard-worked items (see $\bar{f}$ in Table 8). Moreover, to judge from the fractions of simplex formations to the total numbers of types, verbs (0.52) do less well than nouns (0.43) or adjectives (0.31). This suggests that verbs are the basic items of the language, and that it is relatively difficult to extend the set of verbs. Second, the formation of verbs may be a more complex process than the formation of nouns or adjectives, for reasons pertaining to the establishing of predicate-argument structure. The relative complexity of verb formation may be a factor that limits the degree of productivity of complex verbs.

Ultimately, then, the best way to proceed when analyzing morphological productivity quantitatively is to investigate the shape of the whole frequency distribution. In many cases, $\mathscr{P}$ will adequately summarize the properties of the distribution that are of interest, but deviations at the head and tail of the distribution may require additional scrutiny if the productivity of associated WFR is to be properly understood from the quantitative point of view.

4. Global productivity revisited

In this section we briefly discuss a parametric technique which can be used to assess the global productivity P^* of WFRs more precisely than is possible on the basis of $\mathscr{P}$ and V. The basic idea is that if we are able to obtain an estimate of the possible number of types a WFR can give rise to, that is, the number of types S in the population, we are in a better position to evaluate P^*.

In order to obtain an estimate of S, we have to enrich the statistical model on which $\mathscr{P}$ is based with additional assumptions concerning the shape of the frequency distribution. One such assumption is that the number of types occurring once should be greater than the number of types occurring twice, and that the same inequality should hold for the numbers of types occurring two and three times, three and four times, etc. ($n_r > n_{r+1}$). Note that this immediately rules out the application of this theory to frequency distributions for which the number of types occurring twice is greater than the number of hapaxes, such as the frequency distribution of *un-* (see Table 4). Another such assumption is that the number of types that occur with frequency r can be expressed as a function of r. On the basis of these assumptions, and by making use of the information contained in the full frequency distribution rather than using only n_1, V, and N, an estimate of S can be obtained by considering what happens to V when N becomes infinitely large.

We may pause here to consider what happens to $\mathcal{P}=\frac{n_1}{N}$ when $N\rightarrow\infty$. Surprizingly, for ever-increasing sample size N our statistic $\mathcal{P}$ will eventually become zero, even for productive affixes. More formally,

(3) $$\lim_{N\rightarrow\infty}\frac{n_1}{N}=0.$$

This need not be due to the absence of hapaxes. The problem is that N becomes so large that the contribution of the hapaxes is no longer felt. However, a defining characteristic of a fully productive WFR is that the proportion of hapaxes will remain nonnegligeable, that is, that the inequality

(4) $$\lim_{N\rightarrow\infty}\frac{n_1}{V}=0.$$

should hold (see Baayen 1989). (For very large corpora, say ten times the size of the Cobuild corpus, it may turn out that the statistic $\frac{n_1}{V}$ yields more interesting results than our productivity measure $\mathcal{P}$. The reason for preferring $\mathcal{P}$ is, of course, that only $\mathcal{P}$ has a statistical interpretation that is in line with the linguistic notion of productivity.) The linguistic interpretation of all this is that when one has brought together all and any words ever used within a given language at a particular stage in its historical development, one will find that the rate at which new complex formations are added is zero for both productive and unproductive WFRs, but that only fully productive WFRs show up with distributions with hapaxes.

Having discussed the idea of considering a frequency distribution in its limiting form for $N\rightarrow\infty$, we now turn to the estimation of S. A model that is easy to apply is the so-called Waring–Herdan–Muller model (Muller 1979). According to this model, the number of types n_r that occur with token frequency r in the sample can be expressed as a function of r and two additional parameters, a and x. (When a is fixed at zero and x at unity, the Waring–Herdan–Muller model reduces to Zipf's law.) When we use this model to calculate S, the results listed in Table 10 are obtained. It should be stressed that the Waring–Herdan–Muller model is somewhat unreliable, with the resulting effect that it consistently tends to underestimate S. In fact, the Waring–Herdan–Muller model, which has the advantage of computational simplicity, has the serious defect that it rules out a priori the possibility that the underlying population is infinite, thereby introducing a negative bias in the evaluation of S for productive affixes. With this caveat in mind, Table 10 illustrates the fact that, as the numbers of rare events in the population increase, the number of types not occurring in the sample, n_0, and hence S, will also increase.[19]

In addition to S, we have also listed the values obtained for a measure for global productivity that we have found to be useful, namely

(5) $$\mathscr{I} = \frac{V}{S}$$

This index of productivity, the inverse of Aronoff's index of productivity mentioned in the introduction, expresses the extent to which the number of types in the population S exceeds the number of types in the sample V. The WFRs in Table 10 are listed in order of decreasing $\mathscr{I}$, which again provides a productivity ranking that is intuitively more or less in harmony with our intuitive judgments. We may point out that flectional *-er* is now ranked above *-ness* rather than below *-ish*, the ranking obtained on the basis of $\mathscr{P}$: the Waring–Herdan–Muller model is less sensitive to the presence of extremely high-frequency outliers which are in part responsible for the low $\mathscr{P}$ value of flectional *-er*.

5. Summary

We have applied the mathematically motivated measure of morphological productivity $\mathscr{P}$ to the frequency distributions of a large number of English derivational affixes. Apart from the general finding that this measure succeeds in providing a satisfying ranking of affixal classes according to

Table 10. *Indices of productivity according to the Waring–Herdan–Muller model*

18,000,000	V	$\hat{S}$	$\mathscr{I}$	a	x
-er (comp.)	354	489	1.38	2.69	3.73
-ness	497	604	1.22	5.49	6.68
-s (plural)	9728	11435	1.16	6.33	7.44
N-ish	67	77	1.15	9.88	11.35
-ity	405	439	1.08	13.83	14.97
-ation	678	732	1.08	14.31	15.45
-ous	264	278	1.05	24.27	25.58
-ment	184	194	1.05	20.15	21.24

their degree of productivity, a fact which we think is of interest by itself, our study has revealed a number of other points of interest. First, we have uncovered the fact that the category of the base word (nominal versus verbal) has an unexpectedly strong effect on the degree of productivity of derived words. Second, research concerning the effect of restrictions on word-formation rules on the productivity of these rules has up to now been carried out solely on the basis of type counts. We have shown that such counts can be misleading: in the case of abstracta coined on the basis of adjectives in *-ous*, linguistic theory predicts transparent *-ness* to be more productive than nontransparant *-ity*. A count of types suggests the reverse. Nevertheless, an examination by means of $\mathscr{P}$ of the frequency distributions involved reveals the theoretical prediction to be correct. Third, the methodology developed here enables us to make a principled distinction between truly productive 'vogue' morphemes such as *-ee* and unproductive affixes like *-esque* which are sporadically found in neologisms. Finally, we hope to have shown that the gap between theoretical linguistics on the one hand and quantitative linguistics on the other is not unbridgeable, and that the interaction of the two approaches aids the goal of understanding the phenomenon of human language.

Received 22 January 1991
Revised version received
10 June 1991

Free University, Amsterdam
University of New Hampshire

Notes

* The present address of the first author is the Max-Planck-Institut für Psycholinguistik, Wundtlaan 1, 6525XD Nijmegen, The Netherlands. The second author wishes to acknowledge the generous support of NWO, the Netherlands Scientific Organization, during the time that this paper was written.

1. See Lieber (i.p.) for an extensive application of this measure to a theory of morphology.
2. In the Celex release underlying the present discussion, the derivational parsing was not yet made available. Consequently, the data presented in this paper are based on our own morphological analyses and hence may differ slightly from the Celex parsing of later releases.
3. We follow Good and Toulmin (1956) and Efron and Thisted (1976) when we use the notations *V* ('vocabulary') and *S* ('species') to refer to the random variables of the numbers of types in sample and population respectively.
4. Aronoff (1982) and Anshen and Aronoff (1988) calculate the mean token frequencies on the basis of the token frequencies of the types occurring in the Kučera and Francis (1967) corpus, all of which have a frequency greater than unity, combined with the zero frequencies of those types which do not occur in the corpus but which are listed in Walker (1936). Since the number of such zero frequency formations is much larger for *-iveness* than for *-ivity*, this procedure has the effect of lowering the mean frequency

of *-iveness* to a much larger extent than for *-ivity*. For the data of Table 1, this procedure would result in a mean frequency of 3.47 for *-iveness* and 160.52 for *-ivity*, raising the ratio of the means of *-ivity* and *-iveness* from 11.91 to 46.25, suggesting that token frequency is far more important for *-ity* than for *-ness*. Unfortunately, this way of calculating and comparing means is illegitimate. First, as mentioned above, dictionaries are a questionable source for gaining insight into the range of use of the types. Second, while data from different sources can be compared and subjected to statistical analysis, the mixing of frequency data from a corpus with data from a dictionary is highly questionable, since it is entirely unclear on what kind of sample space our probability measure has to be defined. Either we use the actual corpus data, as in Table 10, or we restrict ourselves to comparing types in corpus and dictionary.

5. For instance, word frequency has been found to correlate positively with the number of dictionary meanings (Reder et al. 1974), and in elicitation high-frequency words yield larger numbers of different associative responses than low-frequency words (Paivio et al. 1968).
6. See also Rainer (1988), who argues convincingly that the competence-theory notion of listing does not carry over to theories of the mental lexicon.
7. The kind of semantic sensitivity at stake here differs from the kind of semantic factors that can be analyzed as conditions on WFRs. True semantic conditions on WFRs, such as Zimmer's observation that *un-* tends not to attach to semantically negative base words (**unsad*), concern semantically incongruous combinations of base and affix. In contrast, Thorndike's ratios are sensitive to lexical areas where derived formations have a higher or lower extent of use: *redly* is semantically felicitous, even though a context in which it could be used is somewhat difficult to imagine.
8. For a mathematically more precise definition and discussion of the measure $\mathscr{P}$ the reader is referred to Baayen (1989: chapter 5).
9. In the Dutch Eindhoven corpus (600,000 tokens), the $\mathscr{P}$ value of compounds assumes the very high value of 0.225 for the substantial number of 4,277 types.
10. The best way to proceed when comparisons are made across corpora of different size, say C_1 and C_2, $C_1 > C_2$, is to sample at random C_2 tokens from the larger sample and to select the individual samples of tokens with a particular morphological property from this reduced corpus.
11. The figures in Table 3 reflect the occurrence of both those words in *-ation* whose base actually contains *-ate* (like *complicate*), and those where *-at* is clearly part of the suffix (like *representation*). Figures for the variants *-ution* (*resolution*), *-ition* (*competition*), and *-ion* (*rebellion*, *abstraction*) were calculated separately. All of these allomorphs have a P value somewhat lower than that of *-ation*.
12. Within some theories of generative morphology (such as Pesetsky 1985; Lieber 1989, i.p.) prefixes such as *un-*, *in-*, and *re-* are treated as categoryless affixes. The bases to which they attach supplies the category of the derived word through a process of percolation.
13. The three words in *-esque* in the database are *grotesque* (142), *picaresque* (4), and *picturesque* (92). Only the last formation is a regular denominal adjective, so that, strictly speaking, we are dealing with a single high-frequency word for which it is extremely unsatisfying to assume that it is associated with a productive rule.
14. Anshen and Aronoff made use of 60 subjects in all, 30 for *-ibility* and *-iveness* and 30 for *-ibleness* and *-ivity*.
15. Since the modes of the frequency distributions of the simplex nouns and the simplex adjectives do not lie at n_1 (for the simplex nouns at n, for the simplex adjectives at n_5), the methods suggested in Good and Toulmin (1956) for estimating the variance

of n_1 cannot be applied. Hence significance testing on the basis of $\mathscr{P}$, as suggested in Baayen (1989), cannot be carried out. Although somewhat impressionistic, the differences observed are generally large and, in the light of the large numbers of observations involved, strongly suggest marked differences in the frequency distributions.

16. According to Booij (1988), *-er* binds the external argument of the verbs to which it attaches and therefore must attach to verbs which have external arguments; agentive/instrumentals from verbs like *arrive* and *seem* are therefore impossible (**arriver*, **seemer*).
17. *Re-*, like *un-* and *in-*, is not a category-bearing prefix according to Pesetsky (1985) and Lieber (1989, i.p.). We treat it here since the category of the derived word is verb, even though it is the base, rather than the prefix, according to these theories, which supplies the category of the resulting word.
18. For instance, the use of *re-* in its iterative sense may well suffer severe quantitative inhibition by the availability of a very simple syntactic strategy for the expression of iteration, namely the use of the adverb 'again'.
19. Better results can be obtained with Chitašvili and Khmaladze's (1989) extended generalized Zipf's law. The programs for estimating the parameters of this model are at present being developed, which is the reason why this more accurate model has not been used here. Other techniques, both nonparametric (Efron and Thisted 1976) and parametric (Sichel 1975; Carroll 1969), are also available.

References

Anshen, Frank, and Aronoff, Mark (1988). Producing morphologically complex words. *Linguistics* 26, 641–655.

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

—(1982). Potential words, actual words, productivity and frequency. *Preprints of the Plenary Session Papers of the XIIIth International Congress of Linguists*, 141–148. Tokyo.

Baayen, R. Harald (1989). A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation. Unpublished dissertation, Vrije Universiteit, Amsterdam.

—(forthcoming). Quantitative aspects of morphological productivity. *Yearbook of Morphology*.

Bauer, Leslie (1983). *English Word Formation*. Cambridge: Cambridge University Press.

Bolinger, Dwight L. (1948). On defining the morpheme. In *Forms of English. Accent, Morpheme, Order*, D. L. Bolinger (ed.), 183–189. Cambridge, MA: Harvard University Press.

Booij, Geert E. (1988). The relation between inheritance and argument linking: deverbal nouns in Dutch. In *Morphology and Modularity*, M. Everaert, A. Evers, and M. Trommelen (eds.), 57–74. Dordrecht: Foris.

Carroll, John B. (1967). On sampling from a lognormal model of word frequency distribution. In *Computational Analysis of Present-Day American English*, H. Kučera and W. N. Francis (eds.), 406–424. Providence: Brown University Press.

—(1969). *A Rationale for an Asymptotic Lognormal Form of Word Frequency Distributions*. Research Bulletin, Educational Testing Service. Princeton: Educational Testing Service.

Chitašvili, Reso J., and Khmaladze, Estate V. (1989). Statistical analysis of large number

of rare events and related problems. *Transactions of the Tbilisi Mathematical Institute* 91, 196–245.
Cutler, Ann (1980). Productivity in word formation. *Chicago Linguistic Society* 16, 45–51.
Downing, P. (1977). On the creation and use of English compound nouns. *Language* 53, 810–842.
Efron, Bradley, and Thisted, Ronald (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* 63, 435–447.
Good, I. J., and Toulmin, G. H. (1956). The number of new species and the increase in population coverage, when a sample is increased. *Biometrika* 43, 45–63.
Harwood, F. W., and Wright, A. M. (1956). Statistical study of English word formation. *Language* 32, 260–273.
Kastovsky, Dieter (1986). Productivity in word formation. *Linguistics* 24, 585–600.
Kučera, Henry, and Francis, W. Nelson (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.
Levin, Beth, and Rappaport, Malka (1986). The formation of adjectival passives. *Linguistic Inquiry* 17, 623–662.
Lieber, Rochelle (1989). On percolation. *Yearbook of Morphology* 2, 95–138.
— (i.p.). *Deconstructing Morphology: Word Formation in Syntactic Theory*. Chicago: University of Chicago Press.
Lyons, John (1977). *Semantics*. Cambridge: Cambridge University Press.
Marchand, Hans (1969). *The Categories and Types of Present-Day English Word-Formation*. Munich: Beck.
Muller, Charles (1979). Du nouveau sur les distributions lexicales: la formule de Waring–Herdan. In *Langue Française et Linguistique Quantitative*, Charles Muller, 177–195. Geneva: Slatkine.
Paivio, Allan, Yuille, John C., and Madigan, Stephan A. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Joural of Experimental Psychology Monograph* 76 (1), part 2.
Pesetsky, David (1985). Morphology and logical form. *Linguistic Inquiry* 16, 193–248.
Rainer, Franz (1988). Towards a theory of blocking: the case of Italian and German quality nouns. *Yearbook of Morphology* 1, 155–185.
Rappaport, Malka, and Levin, Beth (1986). The formation of adjectival passives. *Linguistic Inquiry* 17, 623–662.
—, and Levin, Beth (1988). What to do with θ-roles. In *Syntax and Semantics*, vol. 21, W. Wilkins (ed.), 7–36. New York: Academic Press.
Reder, Lynne M., Anderson, John R., and Bjork, Robert A. (1974). A semantic interpretation of encoding specificity. *Journal of Experimental Psychology* 102, 648–656.
Renouf, Antoine (1987). Corpus development. In *Looking Up: An Account of the Cobuild Project in Lexical Computing*, J. M. Sinclair (ed.), 1–40. London: Collins.
Scalise, Sergio (1988). Inflection and derivation. *Linguistics* 26, 561–581.
Schultink, Henk (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren* 2, 110–125.
—(1962). *De morfologische valentie van het ongelede adjectief in modern Nederlands*. The Hague: van Goor.
Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association* 70, 542–547.
Thorndike, E. L. (1943). Derivation ratios. *Language* 19, 27–37.
van Dale (1976). *Groot Woordenboek der Nederlandse Taal*, 10th ed. C. Kruyskamp (ed.). The Hague: Nijhoff.

van Marle, Jaap (1985). *On the Paradigmatic Dimension of Morphological Creativity*. Dordrecht: Foris.

Walker, J. (1936). *Walker's Rhyming Dictionary*. New York: Dutton.

Williams, Edwin (1981). Argument structure and morphology. *Linguistic Review* 1, 81–114.

Zimmer, Karl E. (1964). *Affixal Negation in English and Other Languages*. *Word* 20, supplement.