Title
Defining New Words in Corpus Data: Productivity of English Suffixes in the British National Corpus

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 26(26)

ISSN
1069-7977

Author
Nishimoto, Eiji

Publication Date
2004

Peer reviewed

Defining New Words in Corpus Data: Productivity of English Suffixes in the British National Corpus

Eiji Nishimoto (enishimoto@gc.cuny.edu)
Ph.D. Program in Linguistics, The Graduate Center
The City University of New York
365 Fifth Avenue, New York, NY 10016 USA

Abstract

The present study introduces a method of identifying potentially new words in a large corpus of texts, and assesses the morphological productivity of 12 English suffixes, based on some 78 million words of the written component (books and periodicals) of the British National Corpus (BNC). The method compares two corpus segments (created by randomly sampling at the level of documents within the BNC), and defines new words as those that are not shared across segments (segments being interpreted as randomly sampled speaker groups). The approach taken differs from others in the literature in that new words are identified irrespective of how many times a given word is used by the same speaker (author). A productivity ranking of the 12 English suffixes is obtained, and the results are shown to be intuitively satisfying and stable over different sample sizes. With a psycholinguistic interpretation of the data, implications for the nature of intuitions about productivity are considered.

Introduction

Morphological productivity is central to the study of word formation, but it continues to defy a solid, uniform description (see e.g., Aronoff, 1976; Bauer, 2001; Plag, 1999). The coinage of a "new" word is abundant in our daily use of language; for example, a person who is being gossiped about may be referred to as a *gossipee*, or a used book may be *cleanish*. Affixation in English (as in *gossip* + *-ee* → *gossipee*; *clean* + *-ish* → *cleanish*) is a productive word formation process, and there is plenty of evidence that affixes differ in their *degree of productivity* (e.g., Aronoff, 1976; Bauer, 2001); for example, words can in general be formed more easily with *-ness* than with *-ity* (and thus we may accept *cleanness* but not *cleanity*). The majority of researchers investigating the issue of productivity are interested in accounting for varying degrees of productivity, and several productivity measures have been proposed in the literature (e.g., Aronoff, 1976; Baayen, 1992, 2001; Bauer, 2001; Plag, 1999). Assessing the degree of productivity, however, has proven to be a complex task (Bauer, 2001): while the consensus seems to be that capturing the coinage of new words is essential in assessing productivity, there is an inherent difficulty in defining what a "new" word is.

Most notable among previous studies is a corpus-based approach proposed by Baayen (1992, 2001). Based on word frequency in a large corpus of texts, his productivity measure is formulated as $P = n_1/N$, where given a particular affix, n_1 is the number of word types with that affix that occur only once (the so-called *hapax legomena*, hereafter *hapaxes*), N is the sum of word tokens with that affix, and P is the productivity index.[1] P is interpreted as expressing the probability of encountering a word type with a given affix that has not been seen in the sampled corpus. Thus, new words are defined under this measure as "unseen" words in a corpus. An important characteristic of P is that it is based on token frequency—N directly refers to a count over tokens, and a word is included in the n_1 count only if it occurs just once. The measure P, with its focus on hapaxes as estimators of unseen words, is motivated by the probability estimation method of Good (1953)—or the *Good-Turing* estimation method (Church & Gale, 1991).[2]

While a dictionary provides another source of data for quantifying morphological productivity, a corpus-based approach has many advantages. A large corpus of texts contains productively formed words that are typically not listed in a dictionary (e.g., *gossipee*), and corpus data reflect how words are actually used (Baayen & Lieber, 1991; Baayen & Renouf, 1996).

The present study pursues and extends the corpus-based approach by introducing a new method of identifying new words and assessing productivity.

Type Frequency and Deleted Estimation

It has been suggested that the type frequency for an affix (the number of word types with an affix) in a corpus, represented by V, is inadequate in expressing its degree of productivity. Baayen and Lieber (1991: 804) point out that in their reference corpus of 18 million words, the type frequencies for *-ness* (497) and *-ity* (405) do not adequately express the fact that *-ness* is intuitively felt to be much more productive than *-ity*. They find that the P indices for *-ness* (0.0044) and *-ity* (0.0007) are more in line with linguists' intuitive estimates for these suffixes. There are, however, some aspects of the measure P that can be quite counter-intuitive. In Baayen and Lieber (1991), for example, the P index for verbal suffix *-ize* (0.00007) is substantially lower

[1] As is usually the case in a corpus study, the term *token* refers to each occurrence of a word, and the term *type* refers to each distinct word. For instance, if we have {*awareness*, *fairness*, *fairness*, *sharpness*, *sharpness*}, the *token frequency* for *-ness* is 5 (the sum of all occurrences of *-ness*), whereas the *type frequency* for *-ness* is 3 (the number of distinct words with *-ness*).

[2] For more detail, see Baayen (2001).

than that for *-ity* (0.0007). To correctly interpret these data, we need to take into account the fact that P is dependent on token frequency, and that verbs and nouns generally differ in their overall frequency in a corpus (Baayen & Lieber, 1991). Consequently, an across-the-board comparison of affixes across lexical categories is ruled out.

The view that type frequency in a corpus is problematic for assessments of degree of productivity holds only if type frequency alone is examined, for an entire corpus. A use of type frequency is suggested by Nishimoto (2003) in a productivity measure that adopts the mechanism of *deleted estimation* (Jelinek & Mercer, 1985; see also Manning & Schütze, 1999: 210–211), a probability estimation method used in Language Technology. The basic concept underlying the proposed productivity measure is the cross-comparison of corpus segments to identify word types that are not shared. The P_{DE} measure, a productivity measure based on the deleted estimation method, is formulated as:

$$(1)\quad P_{\mathrm{DE}} = \frac{V_0^{\mathrm{AB}} + V_0^{\mathrm{BA}}}{V^{\mathrm{A}} + V^{\mathrm{B}}} = \frac{(V_0^{\mathrm{AB}} + V_0^{\mathrm{BA}})/2}{(V^{\mathrm{A}} + V^{\mathrm{B}})/2} = \frac{V_{\mathrm{N}}}{V}$$

Given a particular affix and two corpus segments A and B (both of some size m), V^{A} is the number of word types with that affix that are present in segment A, and V_0^{AB} is the number of word types with that affix that are present in segment A but are absent (unseen) in segment B. V^{B} and V_0^{BA} are defined similarly. Averaging the elements of the denominator and of the numerator separately, we obtain V, the total number of word types with that affix in a corpus segment of size m, and V_{N} (V-New), the number of *otherwise-unseen* word types with that affix (unseen being dependent on the relationship between segments). P_{DE} expresses the degree of productivity of an affix as the likelihood that a given word type with an affix will be unseen, hence potentially new. In addition to V and V_{N}, we also define V_{NN} (V-Non-New) as the number of word types with the relevant affix that are seen in both segments, hence non-new. What is essentially achieved by the P_{DE} measure is the division of sampled word types (V) into new word types (V_{N}) and non-new word types (V_{NN}). The relationship $V = V_{\mathrm{N}} + V_{\mathrm{NN}}$ holds in each application of the measure.

What are the grounds for associating new words with words that are not shared by two corpus segments? In the *British National Corpus* (BNC), data from unique sources are sampled in single documents, and thus each document could be considered to represent a set of words used either by one speaker (author) or by a few speakers (co-authors). Randomly distributing these documents into two corpus segments therefore gives us two groups of randomly sampled speakers (and the words that they used). Words that are not used in common by the two speaker groups are more likely to be new than words used by both groups. Cross-comparing two corpus segments offers a crude yet computationally simple method of separating words into potentially new words and potentially non-new words.[3]

Simulation Environment

We will examine the performance of the P_{DE} measure, based on the written component (books and periodicals) of the BNC, which offers some 78 million words sampled in 2,688 documents. Each sampled document is randomly assigned to one of two corpus segments, until each segment has a specified number of words in total (say, 30 million words). Documents are sampled without replacement, so no document is shared by two corpus segments. One simulation run (i.e., one application of the P_{DE} measure) consists of creating two corpus segments (as above) and obtaining values for V, V_{N}, and P_{DE}, based on formula (1).

Table 1 lists 12 English suffixes and 1 non-suffix control selected for the current study.[4] At least one suffix is included for each major lexical category.

Table 1: 12 English suffixes and 1 non-suffix control.

Suffix	Category	Prediction
-ness, *-ity*	Nominal	*-ness* > *-ity*
-er, *-ee*	Nominal	*-er* > *-ee*
-ion, *-ment*	Nominal	*-ion* > *-ment*
-th	Nominal	Unproductive
-ish, *-ous*	Adjectival	*-ish* > *-ous*
-ize, *-ify*	Verbal	*-ize* > *-ify*
-ly	Adverbial	Productive
ch#	Noun ending	Unproductive

The predicted differences in productivity in the last column of Table 1 are largely based on views expressed in the literature. We also examine *ch#*, the word ending of a noun (as in *church*), as a presumably unproductive non-suffix control that provides a baseline for determining whether suffixes are productive (or unproductive). Different semantic patterns among words formed with a suffix are ignored: for example, *amputee*, *absentee*, and *employee* exhibit different semantic patterns, but they are collectively treated as *-ee* words. We do not distinguish words with a suffix by the class of bases that the suffix attaches to: for example, *-er* includes *employer* (verb base) and *islander* (noun base). Ordinal numbers are excluded from *-th*.

A database of 17,347 word types representing the 12 suffixes and the non-suffix control was compiled, based on 100 million words occurring in the entire BNC. The database crucially relies on decisions about what constitutes a word type with a suffix. Most problematic are prefixation and compounding, which could dramatically increase the number of word types with a suffix. Removing all prefixes

[3] Nishimoto (2004) offers more detailed exploration of the mechanism of the P_{DE} measure, by increasing the number of cross-compared corpus segments (speaker groups) to 6.

[4] These are suffixes whose productivity is often discussed in the morphology literature. We focus on suffixes only, as they play a more prominent role than do prefixes in English word formation.

has some negative consequences, such as *encouragement* → **couragement* or *disagreement* → *agreement*.[5] On the other hand, allowing all prefixes does not seem plausible, since words such as *anti-institution* that appear to be cases of prefixation would count as distinct word types with a suffix. Compounding poses a similar problem, and the issue is further complicated by the variable hyphenation of words. In solving this familiar problem, we make use of entries in the *Oxford English Dictionary* (OED) and *Webster's Third New International Dictionary* (WD). All prefixed forms and compounds are checked against the OED/WD, and any preceding part of a word that cannot be spelled without a hyphen in both the OED and WD is removed. As a result, for example, *anti-institution* will be treated as *institution*, but *disagreement* will remain as *disagreement*. With the assumption that the OED and WD are conservative in accepting novel word forms, the current treatment effectively prevents novel cases of prefixation and compounding from inflating the count of word types with a suffix. Each word type in the database was inspected to exclude errors (e.g., misspelled words, words with a pseudo-suffix).[6] See Nishimoto (2004) for further detail.

Evaluation of Data

Productivity Indices

Table 2 presents mean values for V, V_N, and P_{DE}, averaged over 100 simulation runs, with 30 million words in each of the two corpus segments required by the P_{DE} measure (i.e., the total of 60 million words were sampled in each run). The suffixes in Table 2 are sorted by their P_{DE} value, to achieve a productivity ranking.

Suffixes *-ish* and *-ness* meet our expectations by being found at the more productive end of the ranking (although we might have expected *-ness* to be more productive than *-ish*), and *-th* and *ch#* fall at the less productive end. We consider the P_{DE} index for *ch#* to arise from processes other than affixation, such as the coinage of simplex words, compounding, or some sources of noise including the occurrence of rare or obsolete words.

Taking *ch#* as a baseline for determining productivity in affixation, we find that *-th* is unproductive. The finding that *-ment* is effectively non-productive matches Bauer's (2001: 8–9) observation that the productivity of *-ment* has been in decline so that new *-ment* words are synchronically rare.

Table 2: Mean values of the P_{DE} measure.

	V	V_N	P_{DE}
-ish	261.3	90.6	0.347
-ness	1354.9	431.2	0.318
-ee	88.6	26.1	0.295
-ize	437.6	114.5	0.262
-ity	1008.5	234.4	0.232
-er	2517.8	558.6	0.222
-ly	3585.0	754.3	0.210
-ify	105.8	21.1	0.199
-ous	639.1	107.1	0.168
-ion	2152.9	348.7	0.162
-ment	424.2	61.6	0.145
ch#	213.6	29.7	0.139
-th	40.9	3.5	0.085

The high productivity of *-ee* is somewhat unexpected, on its face. Based on the measure *P*, Baayen and Lieber (1991) also find *-ee* (0.0016) to be more productive than *-er* (0.0007), and they attribute the high productivity of *-ee* to the "vogue" nature of this suffix, as suggested by Marchand (1969).

In contrast to *-ee*, the productivity of *-er* is lower than we might have expected. Also lower than expected is the P_{DE} index for *-ly*. The result for *-ly* seems unsatisfactory considering the high regularity in *-ly* word formation—the suffix attaches to almost any adjective to form an adverb, with few restrictions (Aronoff, 1976: 37 fn 4; Baayen & Renouf, 1996: 82–83). The low P_{DE} indices for *-er* and *-ly* might be thought to arise from large values of V for these suffixes; however, Spearman's test shows no significant correlation between V and P_{DE}, $r_s = .203$, $p > .10$. We will return later to the data for *-er* and *-ly*.

Overall, we find that the P_{DE} measure yields results that largely accord with the productivity expected for the suffixes examined.

Sample Size Dependency

A question that naturally arises in evaluating the P_{DE} measure is to what extent the measure is dependent on sample size. Could it be the case, for example, that the productivity ranking of suffixes would differ markedly if the two corpus segments were smaller? Table 3 presents P_{DE} as a function of corpus segment size.[7] Again, each P_{DE} value is a mean over 100 simulation runs.

We find that P_{DE} values are remarkably similar across three series with different corpus-segment sizes. Friedman's test finds no significant difference in P_{DE} among the three, $\chi^2(2,13) = 2.627$, $p > .10$. Spearman's test shows that the P_{DE} indices are highly positively correlated: for 10 vs. 20 million words, $r_s = .990$, $p < .01$; for 10 vs. 30 million words, $r_s = .971$, $p < .01$; and for 20 vs. 30 million words, $r_s = .984$, $p < .01$. Thus, the P_{DE} measure offers a consistent

[5] Removing *dis-* from *disagreement* appears to be undesirable if we view *disagreement* as a nominalization of *disagree*.

[6] There are 6,797 rules defined for these corrections (mostly generated automatically, but some inevitably defined manually for cases such as "dona-a-a-ation" → *donation*). The number of rules is large, but it must be noted that some are needed to obtain correct forms for irrelevant words (so that they can be deemed irrelevant), and that a given word can be misspelled in a number of ways. Errors in a corpus cannot be overlooked. Evert and Lüdeling (2001) point out, for example, that each error in a corpus typically occurs only once and could greatly distort the number of hapaxes.

[7] As is clear from the formulation of the measure, a change in corpus segment size applies simultaneously to both corpus segments.

characterization of the productivity of these suffixes over different sample sizes.

Table 3: P_{DE} as a function of corpus segment size.

	P_{DE}		
	10 million	20 million	30 million
-ish	0.322	0.332	0.347
-ness	0.371	0.336	0.318
-ee	0.313	0.301	0.295
-ize	0.262	0.260	0.262
-ity	0.238	0.235	0.232
-er	0.260	0.236	0.222
-ly	0.226	0.215	0.210
-ify	0.179	0.191	0.199
-ous	0.164	0.165	0.168
-ion	0.169	0.163	0.162
-ment	0.153	0.148	0.145
ch#	0.164	0.153	0.139
-th	0.078	0.081	0.085

Token Frequency of New Words

One advantage of the P_{DE} measure is that new words in a corpus are identified in a way that is not solely dependent on token frequency. To ensure that advantage, it is crucial to implement the measure by creating corpus segments via random sampling at the level of documents (hereafter *RD*), rather than random sampling at the level of words (hereafter *RW*). The data presented in the preceding sections arise in implementations using RD.

If we were to follow RW, which words become identified as new would be dependent on their token frequency in the whole corpus.[8] Under the P_{DE} measure, a word is identified as new if all its tokens are distributed into only one corpus segment. If we were to randomly distribute words into two corpus segments (i.e., RW), the probability P that word w with token frequency r (in the whole corpus) will be identified as new is given by: P(w: new) = $2(0.5^r)$. Figure 1 shows how P(w: new) changes as a function of r. We find that words that occur more than a few times in the whole corpus are highly unlikely to be identified as new. Hapaxes are exceptional in that they are guaranteed to be new, regardless of whether RD or RW is adopted. What is of interest regarding the difference between RD and RW is how many *non-hapaxes* are found to be new.

[8] The *whole corpus* here refers to the set of data used to create two corpus segments.

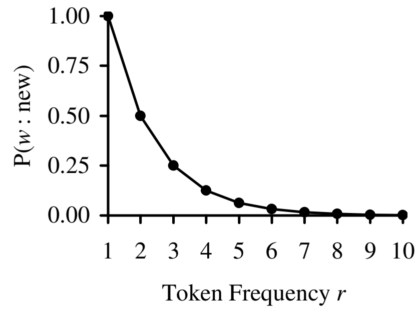

Figure 1: Probability of word w being identified as new.

We compare the outcome of RD and RW as follows. V_N values obtained under RD are listed in Table 2. For each of the 100 simulation runs generating these values, we sum the two corpus segments to obtain the whole corpus, and then, based on token frequencies in this whole, calculate a V_N value expected under RW, $E(V_N)$, based on the following formula:

(2) $$E(V_N) = \sum_{r=1} N_r(0.5^r)$$

Here, r is the token frequency of a word, and N_r is the number of word types that occur r times. Table 4 contrasts V_N under RD and $E(V_N)$ under RW.

Table 4: Mean values for V_N (RD) and $E(V_N)$ (RW).

	V_N	$E(V_N)$
-ish	90.6	85.9
-ness	431.2	401.4
-ee	26.1	18.3
-ize	114.5	101.4
-ity	234.4	186.5
-er	558.6	463.9
-ly	754.3	707.0
-ify	21.1	18.5
-ous	107.1	89.1
-ion	348.7	270.6
-ment	61.6	49.4
ch#	29.7	23.0
-th	3.5	2.8

We find that each value of $E(V_N)$ is consistently an underestimation of the V_N. That is, more new words are captured by RD than by RW.

What kind of words are responsible for the discrepancy between RD and RW that is exhibited in Table 4? Consider *causee* (undoubtedly new to the majority of English speakers, except perhaps those who are syntacticians), which occurs 9 times in 1 document of the written component of the BNC. Under RW, the probability that all 9 tokens of *causee* will be distributed into only a single corpus segment is as low as about 0.002—in effect, *causee* is virtually guaranteed to be identified as non-new under RW.

Under RD, on the other hand, all 9 tokens of *causee* (occurring in just 1 document) will inevitably be distributed into only one corpus segment, and *causee* will thus be identified as new. The advantage of the P_{DE} measure (when implemented with RD) is that it captures as new words those words such as *causee* that are repeatedly used by the same speaker.

Implications for Linguistic Intuitions

Intuition-Based Interpretation

Although intuitions about productivity may not be reliable, they play an informal yet important role in evaluating results for a productivity measure—such results are often said to be intuitive or counter-intuitive. However, to the extent that little is known about the nature of such intuitions, determining the validity of a productivity measure on this basis may not be viable. Nevertheless, we may still ask what kind of information could be available to speakers (linguists) when they offer intuitive judgments about productivity.

We found in Table 2 that P_{DE} indices for *-er* and *-ly* are unexpectedly low, but one possibility is that the data on which P_{DE} is built are simply not in the form to be accessible to intuition. Speakers presumably cannot tell with any precision, for example, how many *-er* and *-ee* word types exist in the BNC, and thus, exact values of V, V_N, and P_{DE} may have little relevance to speakers' intuitions. On the other hand, speakers may be able to predict that "more" *-er* words than *-ee* words will occur.

What will be attempted here is a transformation of the data underlying the P_{DE} measure into a form that could be relevant to speakers' intuitions. There are two points to consider. The first is the possibility that whatever type frequency information speakers may have access to may be better represented on a logarithmic scale. Word frequency effects have been well studied in psycholinguistics (since Howes & Solomon, 1951; see Monsell, 1991, for an overview), where it has been noted that reaction time for a word in a lexical decision task is inversely proportional to the log frequency of that word. Although word frequency effects are normally discussed with respect to token frequency, the possibility that we entertain is that a similar logarithmic scaling may be also applicable to type frequency information.

The second point to consider is Baayen's (1993: 204) view that speakers' intuitive judgments on productivity are ordinal rather than interval in nature. Speakers presumably cannot tell to what extent *-ness* is more productive than *-ity*, but they may reject a productivity ranking in which *-ness* is ranked lower than *-ity*. Baayen also suggests that intuitions about productivity may simply be unavailable for some affixes.

Transforming V and V_N into $\log_{10}V$ and $\log_{10}V_N$ and taking their ratio (by analogy to V_N/V) is too simplistic a solution, and suffers a problem in that the complementary relationship between V_N and V_{NN} will be broken: the order of suffixes defined by V_N/V should be the reverse of the order defined by V_{NN}/V, but that relationship will no longer hold when values are log-transformed values. A solution to this dilemma is to shift our point of view, and to think of $\log_{10}V_N$ as the *extent* to which words are new and of $\log_{10}V_{NN}$ as the *extent* to which words are non-new. These are two conflicting factors that may simultaneously affect speakers' "impression" about a given word formation process. When $\log_{10}V_N$ (the extent to which words with a suffix are new) approaches $\log_{10}V_{NN}$ (the extent to which words with that suffix are non-new), the word formation process for that suffix may be felt to be productive, with a degree that can be calculated by the ratio of $\log_{10}V_N$ to $\log_{10}V_{NN}$.[9] Table 5 presents a productivity ranking of suffixes calculated in just this way.

Table 5: Intuition-oriented productivity ranking of suffixes.

	$\log_{10}V_N$	$\log_{10}V_{NN}$	Ratio
-ness	2.63	2.97	0.886
-ish	1.96	2.23	0.879
-er	2.75	3.29	0.836
-ly	2.88	3.45	0.835
-ize	2.06	2.51	0.821
-ity	2.37	2.89	0.820
-ee	1.42	1.80	0.789
-ion	2.54	3.26	0.779
-ous	2.03	2.73	0.744
-ment	1.79	2.56	0.699
-ify	1.32	1.93	0.684
ch#	1.47	2.26	0.650
-th	0.54	1.57	0.344

Following the view that intuitive judgments on productivity have an ordinal character, we concentrate only on the ranking of suffixes shown in Table 5. Interestingly, we seem to have gained many improvements as compared to Table 2. We particularly note the following: (a) *-ness* now counts as the most productive suffix; (b) *-er* and *-ly* move up in the ranking to be close to *-ness* and *-ish*; (c) *-ee* moves down in the ranking but is still close to *-ize*; and (d) *-ify* is now much lower in the ranking. Perhaps one unsatisfactory result is that *-ly* still does not emerge as the most productive suffix.

Although the exploration offered in this final section is based on speculation about what information could be available to speakers, the fact that the productivity ranking of suffixes in Table 5 is intuitively satisfying, by and large, suggests that the approach merits further investigation in future research.

Conclusion

The analysis of the data for the P_{DE} measure demonstrates that the deleted estimation method offers an effective means of capturing new words in corpus data and of assessing the

[9] The complementary relationship between V_N and V_{NN} is of course maintained by the ratio of $\log_{10}V_{NN}$ to $\log_{10}V_N$.

productivity of affixes. An interesting characteristic of the P_{DE} measure is that its identification of new words is not dependent on token frequency, and this may be construed as an advantage, given potential burstiness in the use of new coinages. The current measure identifies a word as new regardless of whether it is used repeatedly by the same speaker. The measure is also shown to be stable over different sample sizes.

Some findings appeared to deviate slightly from our intuitive expectations, but we proposed, (appealing to a psycholinguistic interpretation of the data), that it may be necessary to draw a distinction between raw corpus statistics and information that could be accessible to intuitions about productivity. Corpus statistics, scaled in psychologically plausible ways, may offer insights into the kind of information available to speakers when they make intuitive judgments on productivity.

A description of productivity obtained with a corpus-based productivity measure will be useful in many forms of linguistic research, not necessarily limited to the study of word formation. The success of the present study provides another indication that the corpus-based approach to the study of productivity advocated by Baayen (1992, 2001) is worthy of many future extensions.

Acknowledgments

The author wishes to thank Dianne Bradley, Harald Baayen, and the anonymous reviewers for their insightful comments on this paper. Any errors remain the responsibility of the author.

References

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

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

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1992* (pp. 181–208). Dordrecht: Kluwer.

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer.

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

Baayen, R. H., & Renouf, A. (1996). Chronicling the Times: productive lexical innovations in an English newspaper. *Language, 72,* 69–96.

Bauer, L. (2001). *Morphological Productivity.* Cambridge, UK: Cambridge University Press.

British National Corpus (World Edition) [CD-ROM]. (2000). Oxford, UK: Oxford University Computing Services.

Church, K. W., & Gale, W. A. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language, 5,* 19–54.

Evert, S., & Lüdeling, A. (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 167–175). Lancaster, UK.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40,* 237–264.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41,* 401–410.

Jelinek, F., & Mercer, R. (1985). Probability distribution estimation for sparse data. *IBM Technical Disclosure Bulletin, 28,* 2591–2594.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Marchand, H. (1969). *Categories and Types of Present-Day English Word-Formation.* München: Beck.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition* (pp. 148–197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Webster's Third New International Dictionary, Unabridged (Version 2.5) [CD-ROM]. (2000). Springfield, MA: Merriam-Webster.

Nishimoto, E. (2003). Measuring and comparing the productivity of Mandarin Chinese suffixes. *Journal of Computational Linguistics and Chinese Language Processing, 8* (1), 49–76.

Nishimoto, E. (2004). *A Corpus-Based Delimitation of New Words: Cross-Segment Comparison and Morphological Productivity.* Doctoral dissertation, City University of New York.

Oxford English Dictionary (2nd ed., Version 3.0) [CD-ROM]. (2002). Oxford, UK: Oxford University Press.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation.* Berlin: Mouton de Gruyter.