

## A Productivity Ranking of English Suffixes: The Use of a Corpus-based, Deleted Estimation Productivity Measure

The present study reports some preliminary results of a quantitative *morphological productivity measure* applied to the *Wall Street Journal (WSJ)* corpus of 45-million English words (LDC, 1993). Morphologists have long been intrigued by word formation processes (affixation in particular) that vary in their degree of productivity (Aronoff, 1976). Studies over the past decades have identified some restrictions on word formation that cause one affix to be less productive than another. For example, the *Latinate restriction* requires the *base* word to be of Latinate origin, and *-ity* (e.g., ‘purity’) which is subject to this restriction is generally considered less productive than *-ness* (e.g., ‘sharpness’) which is free of the restriction (Aronoff, 1976). The literature, however, has lacked a general method of capturing the *degree of productivity* in real language use. That is, there has been no widely-accepted measure that enables a statement such as “*-ness* is twice as productive as *-ity*” or “*-ish* is between *-ness* and *-ity* in productivity.”

A controversial yet influential approach to measuring productivity was initiated by Baayen (1989). His measure of the degree of productivity is expressed by the formula “ $p = n_1 / N$ ” where, given all occurrences of words with a particular affix in a corpus,  $n_1$  is the number of word types with the affix that appear only once in the corpus, the so-called *hapax legomena*,  $N$  is the sum of the token frequencies of words with the affix, and  $p$  is the productivity index of the affix in question. The  $p$  measure expresses productivity as the probability that a “new” word (or an “unseen” word, a word that has not appeared in a corpus) with a particular affix will be encountered in a corpus. Adopting the probability estimation method of Good (1953), the measure approximates the proportion of new words in a corpus by the proportion of hapax legomena in that corpus—as expressed in the above formula. However, it has been argued (van Marle, 1992) that the  $p$  measure, by itself (Baayen, 1993), can yield some unintuitive results that do not lead to a plausible productivity ranking. For example, applying the measure to a corpus of 18-million English words, Baayen and Lieber (1991) find that *-ee* (e.g., ‘employee’) with  $p = 0.0016$  is more than twice as productive as *-er* (e.g., ‘employer’) with  $p = 0.0007$ , and that *-ize* (e.g., ‘computerize’) with  $p = 0.00007$  is ten times “less” productive than *-ity* with  $p = 0.0007$ . As a token-based measure, the  $p$  measure has the characteristic that a large token frequency  $N$  can (sometimes significantly) lower the  $p$  value. Given the limitations of the  $p$  measure, what is yet to be achieved in the literature is a productivity ranking of affixes where the degree of productivity of affixes of different kinds can be compared on a single scale. The underlying objective of the present research is to extend Baayen’s corpus-based approach by exploring new methods of capturing data in a corpus that are relevant to the description of productivity.

The new productivity measure used in the present study is introduced as the  $P_{ide}$  measure in Nishimoto (in press) where a plausible productivity ranking of some Mandarin Chinese suffixes is obtained. The  $P_{ide}$  measure achieves stable and intuitive results by (1) taking a type-based approach (i.e., it does not rely on word-token frequencies), and (2) using an empirical (as opposed to mathematical) method of finding new/unseen words in corpus data—a method that derives from the notion of *deleted estimation* (Jelinek & Mercer, 1985), or more generally, the notion of *twofold cross validation*. A type-based measure has the potential to produce results that are intuitive and easy to interpret, as word types have traditionally played a central role in the discussion of productivity in the literature. As for empirically defining new/unseen words in a corpus, it is accomplished by having two corpora of the same size and text type: unseen words in

one corpus are identified on the basis of their non-occurrence in the other corpus. The  $P_{ide}$  measure is expressed by the following formula:

*Given Corpus A and Corpus B of the same size and text type, and all word types with a particular affix found in these corpora,*

$$P_{ide}(A, B) = \frac{\text{"unseen word types in A given B"} + \text{"unseen word types in B given A"}}{\text{"all word types in A"} + \text{"all word types in B"}}$$

where *unseen word types* in one corpus are those that are absent in the other corpus, and *all word types* in a corpus are all the word types found in that corpus, and  $P_{ide}$  is the degree of productivity of the affix in question. The measure expresses productivity as the likelihood (averaged between two corpora) that a given word with a particular affix will be of a new, unseen type. A productivity ranking of English suffixes obtained from the WSJ corpus suggests that the proposed measure is promising:

<i>suffix</i>	$P_{ide}$	<i>suffix</i>	$P_{ide}$	<i>suffix</i>	$P_{ide}$
<i>-ness</i>	0.383	<i>-er</i>	0.253	<i>-ity</i>	0.202
<i>-ee</i>	0.288	<i>-able</i>	0.248	<i>-ment</i>	0.142
<i>-ism</i>	0.283	<i>-ize</i>	0.247	<i>-ify</i>	0.141
<i>-ish</i>	0.264	<i>-ian</i>	0.244	<i>-ous</i>	0.122
<i>-ist</i>	0.254	<i>-ly</i>	0.211	<i>-th</i>	0.043

The quantitative description of the productivity of English affixes that the  $P_{ide}$  measure can provide will be of interest not only in morphology but also in other subfields of linguistics where the degree of productivity may be relevant.

### Selected Bibliography

- Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Baayen, H. (1989). *A corpus-based study of morphological productivity: Statistical analysis and psychological interpretation*. Doctoral dissertation, Free University, Amsterdam.
- 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, H., & Lieber, R. (1991). Productivity and English word-formation: A corpus-based study. *Linguistics*, 29, 801-843.
- 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.
- Nishimoto, E. (in press). Measuring and comparing the productivity of Mandarin Chinese suffixes. In Sproat, R. (Ed.), *Word formation and Chinese language processing*. Special Issue of the Journal of Computational Linguistics and Chinese Language Processing.
- Van Marle, J. (1992). The relationship between morphological productivity and frequency: A comment on Baayen's performance-oriented conception of morphological productivity. In G. Booij, & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 151-163). Dordrecht: Kluwer.