Abs-234

Stefan Evert (University of Osnabrück)

Quantitative measures of productivity and their significance

The productivity of type-token distributions is an important empirical quantity for corpus-driven approaches to language and many other subfields of linguistics. In addition to recent work on quantitative notions of morphological productivity (e.g. Baayen 1991, 2001; Lüdeling & Evert 2005), applications range from studies of the type-richness e.g. of an author's vocabulary (Efron & Thisted 1976), over stylometrics and authorship attribution (see Juola 2006 for an overview) to patholinguistics (Garrard et al. 2005).

Surprisingly, though, no solid methodological foundation for quantitative studies of productivity has been developed yet. Various measures were suggested in the literature, including the type-token ratio (TTR), Baayen's (1991) productivity index P (hapax legomena / number of tokens), Aronoff's (1976) productivity index I (observed types / possible types), Zipf's law (where the exponent of the rank-frequency law serves as a measure of the "Zipfianness" and hence productivity of the distribution), and even sophisticated statistical models that generalise from finite samples to the type-richness and Zipfianness of the underlying population (so-called LNRE models, Khmaladze 1987, Baayen 2001).

However, there are three fundamental methodological problems shared by all these approaches:

1. Most quantitative measures depend systematically on sample size (i.e. the size of the corpus for which they are computed). This can easily be demonstrated for TTR and P (as argued e.g. by Evert & Lüdeling 2001), but has also been observed with many of the sophisticated LNRE models (Baayen 2001, Fig. 5.12 on p. 182).

2. Usually, no effort is made to assess the uncertainty due to sampling variation. In particular, it is often unclear whether the difference between two observed productivity values can be deemed significant.

3. The interpretation of most productivity measures remains unclear. What exactly is the quantitative phenomenon underlying our intuitive notion of a productive process? And which measure gives the most accurate representation of this phenomenon?

This paper addresses problems 1. and 2. with the help of simulation experiments and empirical corpus data. Special attention is given to sampling variation of the productivity measures, showing how suitable confidence intervals are obtained and significance tests can be carried out. The empirical study is complemented by a discussion of problem 3., which highlights the different intuitive notions of productivity and type-richness encountered in various fields and relates them to the quantitative measures under consideration.

References

Baayen, R. Harald (1992). Quantitative aspects of morphological productivity. Yearbook of Morphology 1991, pages 109–149.

Baayen, R. Harald (2001). Word Frequency Distributions. Kluwer Academic Publishers, Dordrecht.

Lüdeling, Anke and Evert, Stefan (2005). The emergence of productive non-medical -itis. Corpus evidence and qualitative analysis. In S. Kepser and M. Reis (eds.), Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives, pages 351–370. Mouton de Gruyter, Berlin.

Efron, Bradley and Thisted, Ronald (1976). Estimating the number of unseen species: How many words did Shakespeare know? Biometrika, 63(3), 435–447.

Evert, Stefan and Lüdeling, Anke (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (eds.), Proceedings of the Corpus Linguistics 2001 Conference, pages 167–175, Lancaster. UCREL.

Garrard, Peter; Maloney, Lisa M.; Hodges, John R.; Patterson, Karalyn (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain, 128(2), 250–260.

Juola, Patrick (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233–334.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events. Technical Report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.