Thank you for your abstract, it has been saved with the following information:

Number **180**

Authors

Stefan Evert
FAU Erlangen-Nürnberg
stefan.evert@fau.de
Contact person: Stefan Evert, stefan.evert@fau.de

Submission

Type: Poster

Measures of productivity and lexical diversity

Quantitative measures of productivity and lexical diversity – such as the type-token ratio (TTR), Baayen's (1992) productivity index P, or Yule's (1944) K – play a key role in many corpus studies. They are used to assess the degree of morphological productivity (Baayen 1992; Evert & Lüdeling 2005), to estimate the size of an author's vocabulary (Gani 1975; Efron & Thisted 1976), to investigate stylometric differences between writers and settle questions of disputed authorship (Juola 2006), to study diachronic changes in grammar (Bentz et al. 2014), to assess the readability and difficulty level of a text (Graesser et al. 2004), to explore the linguistic correlates of dementia (Garrard et al. 2005; Le et al. 2011), and as a feature in the multivariate analysis of linguistic variation (Biber 1988).

However, virtually all of the approaches and quantitative analyses found in the literature suffer from a number of serious methodological problems:

1. In most cases, no effort is made to assess the uncertainty introduced by sampling variation (Baroni & Evert 2007) and it remains unclear whether the difference between two observed productivity values can be deemed significant. Even if sophisticated statistical LNRE models (Baayen 2001) are used to compute standard errors, authors fail to take the non-randomness of natural language and the variability of estimated model parameters into account.
2. Most quantitative measures depend systematically on sample size (i.e. the size of the corpus or text 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 sophisticated LNRE models (Baayen 2001, Fig. 5.12 on p. 182). Normalizing all texts to the same length is neither a practicable nor a satisfactory solution.
3. Measures are highly sensitive to the presence of a small number of (lexicalized) high-frequency types, which can have a stronger influence on productivity values than the richness of productively formed types. Measures are also sensitive to noise introduced e.g. by typographical mistakes, OCR errors and lack of spelling normalization. Different editorial conventions as well as choices made by the researcher (e.g. whether to consider lowercase and uppercase spellings as distinct types) can lead to further spurious differences in lexical diversity.
4. Many quantitative measures do not have a clear and obvious linguistic interpretation. This holds in particular for the more sophisticated approaches (such as LNRE models) that account for some of the aforementioned problems; but even for simpler measures, their relation to intuitive notions of productivity and lexical richness remains unclear.

My poster gives an overview of established measures of productivity and lexical richness and discusses the four methodological problems listed above in detail. I will also suggest improved approaches and quantitative measures. All findings will be illustrated with simulation experiments as well as a case study based on a large collection of English literary texts (Capitanu et al. 2016).

References

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

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

Baroni, M. and Evert, S. (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In Proceedings of ACL 2007, 904–911. Prague, Czech

Republic.

Bentz, C.; Kiela, D.; Hill, F.; Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. Corpus Linguistics and Linguistic Theory, 10 (2), 175–211.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Capitanu, B.; Underwood, T.; Organisciak, P.; Cole, T.; Sarol, M. J.; Downie, J. S. (2016). The HathiTrust Research Center extracted feature dataset (1.0). http://dx.doi.org/10.13012/J8X63JT3.

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

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

Gani, J. (1975). Some stochastic models in linguistic analysis. Advances in Applied Probability, 7 (2), 232–234.

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

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36 (2), 193–202.

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

Le, Z.; Lancashire, I.; Hirst, G.; Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary and Linguistic Computing, 26 (4), 435–461.

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

Yule, G. U. (1944). The Statistical Study of Literary Vocabulary. Cambridge: Cambridge University Press.

Note to organizers: