Means of Productivity – On the statistical modelling of the restrictedness of lexico-grammatical patterns

Sascha Diwersy, Stefan Evert, Philipp Heinrich, Thomas Proisl

Since the seminal work on the productivity of morphological patterns carried out by R. Harald Baayen (Baayen & Lieber 1991; Baayen 1992), quantitative measures of productivity have become an important field of study in corpus linguistics, expanding to research on syntactic productivity in recent years (cf. amongst others Barðdal 2008, Zeldes 2012). However, there are still numerous problems and open questions: the almost complete lack of significance tests for differences in productivity, the sample-size dependence of many popular measures; their vulnerability to non-randomness in corpus data; and the difficulty of a meaningful linguistic interpretation of the measures (cf. Evert 2017; Evert et al. 2017). Even sophisticated statistical LNRE models (Baayen 2001; Evert & Baroni 2007) are highly sensitive to non-randomness (Baroni & Evert 2007) and cannot be estimated reliable from small samples (Evert & Pipa 2010).

In this paper we explore the applicability of quantitative productivity measures to the particularly challenging case of lexico-grammatical patterns, which are characterized by small sample sizes and complex type-token distributions (e.g. due to semantic restrictions and to their overlap with fully lexicalized multiword expressions). We propose a new state-of-the-art methodology integrating several recent proposals – notably ECHO-corrected LNRE models (Baroni & Evert 2007), standardized productivity measures (Kubát & Milička 2013), permutation tests (Säily 2011), as well as a combination of cross-validation and bootstrapping approaches (Evert et al. 2017) – and use LNRE-based simulation experiments to assess and improve the interpretability of productivity measures.

In order to illustrate our approach, we will conduct a case study on the use of so called “shell nouns” (Schmid 2000) as subject of copula clauses involving the linking verb *be* and a *that*-clause functioning as subject complement (*e.g.* the noun *fact* in (i) *But the* **fact** *is that the very lack of evidence seems to fan the flames of suspicion* [BNC, CB8: 298]). Based on data extracted from the *British National Corpus*, we will show that the application of productivity measures to lexico-grammatical patterns can be substantially improved by introducing a layer of semantic description (which, in our case, is provided by the classification of shell noun uses established by Schmid (2000:92-291)). Our case study will be completed by a number of functional considerations on the several degrees of restrictedness pertaining to the semantically differentiated lexico-grammatical patterns under investigation and their variation across different genres.

References

Baayen, H.R. (1992). “Quantitative aspects of morphological productivity”, in: Booij, G.E. & van Marle, J. (eds.): *Yearbook of Morphology* 1991, Dordrecht: Kluwer Academic Publishers, 109-149.

Baayen, H.R. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, H.R. & Lieber, R. (1991). “Productivity and English derivation: A corpus-based study”, in: *Linguistics*, 29 (5), 801–844.

Baroni, M. & Evert, S. (2007). “Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling”, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 904–911, Prague, Czech Republic.

Barðdal, J. (2008). *Productivity: Evidence from Case and Argument Structure in Icelandic*. Amsterdam/Philadelphia: John Benjamins.

Evert, S. (2017). “Measures of productivity and lexical diversity”, poster at the *ICAME 2017 Conference*, Prague, Czech Republic.

Evert, S. & Baroni, M. (2007). “zipfR: Word frequency distributions in R”, in: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Posters and Demonstrations Session, 29-32, Prague, Czech Republic.

Evert, S. & Pipa, G. (2010). “Probability estimation of rare events in linguistics and computational neuroscience.” Presentation at *KogWis 2010*, Potsdam, Germany.

Evert, S.; Wankerl, S.; Nöth, E. (2017). “Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch”, in Proceedings of the Corpus Linguistics 2017 Conference, Birmingham, UK.

Kubát, M. & Milička, J. (2013). “Vocabulary richness measure in genres,” in: *Journal of Quantitative Linguistics*, 20(4), 339–349.

Säily, T. (2011). “Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations,” in: *Corpus Linguistics and Linguistic Theory*, 7(1), 119–141.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin e.a. : Mouton de Gruyter.

Zeldes, A. (2012). *Productivity in Argument Selection. From Morphology to Syntax*. Berlin e.a.: Mouton de Gruyter.