

---

Pavel Štichauer  
Charles University in Prague, Czech Republic  
[pavel.stichauer@ff.cuni.cz](mailto:pavel.stichauer@ff.cuni.cz)

## **The quantitative approach to morphological productivity in a diachronic perspective**

### **Introduction: the aim of the paper**

The quantitative corpus-based approach to morphological productivity, based on Baayen's work (1992; 2001; 2008), has become a major paradigm in the synchronic studies of the productivity of word formations processes (e.g. Plag 1999; 2006; Gaeta & Ricca 2002; 2003; 2006; Dal 2003). However, there are still not many studies which would apply the method diachronically, i.e. on diachronic corpora covering, preferably, more periods (cf. Lüdeling & Evert 2005; Scherer 2007; Säily & Suomela 2009; Štichauer 2009). This is obviously due to the fact that, with diachronic corpora, there are more problems to be solved before one can proceed to the calculation of the productivity of a given process (cf., e.g., Baayen 2009:909-910). My aim is to formulate some methodological prerequisites and to show the results on one concrete example.

### **The quantitative notion of morphological productivity**

It is well known that Baayen's corpus-based approach conceives of productivity as the likelihood of observing a new type when sampling a sufficiently large corpus. The gradually increasing number of new types (type frequency,  $V$ ) may in fact be seen as a function of token frequency ( $N$ ): with the increasing number of tokens (given by the corpus size), the number of types will also increase (cfr. Baayen 1992:113). This relation gives rise to the definition of *vocabulary growth curve* and to the notion of *vocabulary growth rate*, the latter being captured by the proportion of *hapax legomena* ( $V_1$ ) to the overall number of tokens. The obvious fact that this measure cannot be used for different-sized corpora can now be easily overcome by two techniques (*binomial interpolation* and *extrapolation* based on LNRE models of word frequency distributions, cfr. Baayen 2001; Evert 2004), which are implemented in the package *zipfR*, recently developed by Marco Baroni and Stefan Evert (Baroni & Evert 2006; <http://zipfr.r-forge.r-project.org/>).

### **Principles of lemmatization**

However, before one can use these techniques, some work of linguistic *pre-processing* is necessary. In diachronic corpora, the process of *lemmatization* is particularly tricky because of two facts. First, the lemmatization requires a particular attention to

orthographic and phonomorphological variants. Second, one needs to eliminate all the types which have nothing to do with the word formation rule in question. If this is not done *manually*, the "extraction noise" (Evert 2005:63) will probably distort all possible outcomes. Furthermore, there are problems with the *inhomogeneity* of the underlying corpus. Diachronic corpora (as opposed to large synchronic corpora) tend to be *inhomogeneous* for at least two reasons: first, there may be different text types with different proportions across the corpora one wishes to compare (cf. Baayen 2009:910); second, there is a strong tendency (perhaps, a stronger one than for synchronic corpora) to the so-called *clustering / repetition effects* (cf. Evert 2005:59). In other words, diachronic corpora tend to display a stronger *non-randomness* than could possibly be corrected for within a statistical model (cf. Evert 2006; Baayen 2009:910).

### **Case study: suffixes *-mento* / *-zione* in Old Italian**

I wish to put forward a partial solution to these problems by presenting one concrete example. I intend to show the diachronic development of two Italian deverbal suffixes *-mento* / *-zione* within the time span that goes from the 13<sup>th</sup> to the 16<sup>th</sup> century. The data have been sampled from four different-sized subcorpora created out of the known corpus LIZ 4.0. All the formations in *-mento* and *-zione* have been lemmatized manually, they have been checked against their particular contexts and they were "filtered" by using five major "type elimination" criteria put forward, among others, by Gaeta & Ricca 2002; 2003; 2006 (*strong opacity, baseless formations, nominal bases / different semantic instruction, derivational inner cycles, specific borrowings*). In order to show the diachronic aspect of the productivity of these two suffixes, the above mentioned tools of lexical statistics, implemented in the package *zipfR*, will be used, especially the technique of extrapolation, as the four subcorpora are of gradually increasing sizes. I will show that the suffix *-mento* tends to be constant in its productivity across the four periods in question, while *-zione* displays interesting diachronic variability.

This diachronic variability is particularly interesting because it has important repercussions on the present-day situation of these two suffixes, as described by Gaeta & Ricca 2002; 2003; 2006. First, it is the diachronically increasing token frequency of *-zione* derivatives. In the present-day Italian, *-zione* is by far the most frequent suffix (cf. Gaeta & Ricca 2003; 2006) and with respect to *-mento* it is, of course, of more limited productivity. Second, it is the proportion of *hapax legomena* to the overall number of types. I wish to show how this quantitative complementarity can be traced back to the 15<sup>th</sup> and 16<sup>th</sup> century where the Latin influence leads probably not to a systematic change in the derivational paradigm (*-mento* vs. *-zione*), but to a massive borrowing process. It is only later that *-zione* settles down as an independent and „available“ (in Corbin's sense) affix, being largely dependent upon other changes in the derivational paradigms (especially the increasing productivity of *-izzare* verbs which are an exclusive input to the *-zione* derivatives (cf. Gaeta & Ricca 2006:70).

### **References**

- Baayen, Harald 1992. Quantitative aspects of morphological productivity. In Booij, Geert, Marle, Jaap van (eds.), *Yearbook of Morphology 1991*, 109-149. Dordrecht: Kluwer.
- Baayen, Harald 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.
- Baayen, Harald 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Baayen, H. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke, Merja,

- Kytö (eds.), *Corpus Linguistics. An International Handbook*, vol. 2, 899- 919, Berlin, Mouton de Gruyter.
- Baroni, Marco and Evert, Stefan 2006. The *zipfR* package for lexical statistics: A tutorial introduction (<http://www.cogsci.uni-osnabrueck.de/~severt/zipfR/>).
- Dal, Georgette 2003. Productivité morphologique: définitions et notions connexes. *Langue française*, 140, 3-23.
- Evert, Stefan 2004. A simple LNRE model for random character sequences. *Proceedings of JADT 2004*, 411-422.
- Evert, Stefan 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2004, published in 2005; available from <http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/>.
- Evert, Stefan 2006. How Random is a Corpus? The Library Metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177-190.
- Gaeta, Livio and Ricca, Davide 2002. Corpora testuali e produttività morfologica: i nomi d'azione in due annate della *Stampa*. In *Parallela IX. Testo – variazione – informatica. Text – Variation – Informatik*, Bauer, R. - Goebel, H. (a cura di), 223-249. Wilhelmsfeld: Gottfried Ebert Verlag..
- Gaeta, Livio and Ricca, Davide 2003. Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data. *Italian Journal of Linguistics/Rivista di Linguistica* 15, 1, 63-98.
- Gaeta, Livio and Ricca, Davide 2006. Productivity in Italian word formation: A variable- corpus approach. *Linguistics* 44, 1, 57-89.
- LIZ 4.0 2001. *Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*, a cura di Pasquale Stoppelli ed Eugenio Picchi. Bologna: Zanichelli.
- Lüdeling, Anke and Evert, Stefan 2005. The Emergence of Non-Medical -itis. Corpus Evidence and Qualitative Analysis. In *Linguistic evidence. Empirical, Theoretical, and Computational Perspectives*, Kepser, S. and Reis, M. (eds.), 315-333. Berlin: Mouton de Gruyter.
- Plag, Ingo 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.
- Plag, Ingo 2006. Productivity. In *The Handbook of English Linguistics*, Aarts, B. and McMahon, A. (eds.), 537-557. Oxford: Blackwell.
- Säily, Tanja and Suomela, Jukka 2009. Comparing type counts: The case of women, men and -ity in early English letters. In *Corpus Linguistics: Refinements and Reassessments* [Language and Computers: Studies in Practical Linguistics 69], Renouf, A. and Kehoe, A. (eds.), 87-109. Amsterdam: Rodopi,
- Scherer, Carmen 2007. The role of productivity in word-formation change. In *Historical Linguistics 2005: Selected papers from the 17th International Conference on Historical Linguistics, Madison, Wisconsin, 31 July - 5 August 2005*, Salmons, Joseph C. and Shannon Dubenion-Smith (eds.), 257-271. Amsterdam: John Benjamins.
- Štichauer, Pavel 2009. Morphological productivity in diachrony: the case of the deverbal nouns in -mento, -zione and -gione in Old Italian from the 13<sup>th</sup> to the 16<sup>th</sup> century. In *Selected Proceedings of the 6th Décebrettes*, Fabio Montermini, Gilles Boyé and Jesse Tseng (eds.), 138-147. Somerville, MA: Cascadilla Proceedings Project. [www.lingref.com](http://www.lingref.com), document #2241.