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 (V₁) 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 13th to the 16th 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 15th and 16th 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, H. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke, Merja,


