Samo-compounds in Russian: a corpus-based quantitative approach to morphological productivity in diachrony

Chiara Naccarato (m.naccarato@studenti.unibg.it)
University of Pavia, University of Bergamo

Third Pavia International Summer School for Indo-European Linguistics – Pavia, 7-12 September 2015

INTRODUCTION

AIM

The aim of the present study is to investigate the productivity of the prefixoid *samo-* ("self") in Russian in a diachronic perspective. This prefixoid, which can be attached to nouns, adjectives and verbs, assigning the meaning of orientation towards oneself (e. g. *samokritika* 'self-criticism') or accomplishment of something autonomously or automatically (e. g. *samovzryvanie* 'self-explosion'), intuitively seems to be productive. However, in order to determine the extent to which this prefixoid could and can be used to create new pertinent forms, it is necessary to perform a quantitative analysis based on a large number of data.

THEORETICAL FRAMEWORK

The approach chosen to study the productivity of the prefixoid *samo-* is corpus-based and quantitative in nature (cfr. Baayen 1992, 1993, 2001, 2008). In a quantitative approach, morphological productivity, i. e. *profitability*, is intended as "the extent to which a morphological process may be employed to create new pertinent forms" (Plag 2006: 539). The measure proposed by Baayen and Lieber (1991) to calculate potential productivity is the ratio between the number of *hapax legomena* (V1) with a given affix and the number of tokens (N) with that affix: P = V1/N. This measure is not adequate to calculate productivity over different-sized corpora, since the value of P is a function of N, hence it depends on the corpus size. However, this measure can be applied if its value is calculated at equal token numbers.

DATA

The empirical basis chosen to carry out this study is represented by the Russian National Corpus, one of the largest resources for the study of Russian. In particular, *samo*-compounds are extracted from four selected subcorpora, which represent four time periods: subcorpus 1 (1700-1799), subcorpus 2 (1800-1899), subcorpus 3 (1900-1999) and subcorpus 4 (2000-2015). The figures (after corpus preprocessing) relative to N (tokens), V (types) and V1 (*hapax legomena*) with the prefixoid *samo-* for each subcorpus and the values of P are presented in Table 1:

Subcorpus	N	V	V1	P
Subcorpus 1	789	51	15	0,019
Subcorpus 2	16,734	209	68	0,004
Subcorpus 3	59,922	563	190	0,003
Subcorpus 4	34,294	551	223	0,006

Table 1. *N (tokens), V (types), V1 (hapax legomena) and P (productivity) for each subcorpus*

The four subcorpora present highly different sizes, so the four periods are unequally represented and the results we get are opposite to what we would expect: the productivity of the prefixoid *samo-* seems to decrease over time (with a slight recovery in the fourth subcorpus). This is due to the fact that the value of P is highly dependent on the sample size.

ANALYSIS

In order to calculate the productivity of the prefixoid *samo-* over different-sized corpora, I resort to parametric statistical models of frequency distribution known as LNRE (*Large Number of Rare Events*) models (Baayen 2001). These models, which allow to estimate the expected values of V and V1 for arbitrary values of N larger than the empirical value of N (through the technique of *extrapolation*), are implemented in the package *zipfR*, a tool for lexical statistics in R (Baroni and Evert 2014). The *Zipf-Mandelbrot* (ZM) model was used to estimate the expected values of V (EV) and V1 (EV1) in the four subcorpora at equal values of N. The results obtained at the equal token value of N=50000 are presented in Table 2:

Subcorpus	EV	EV1	P
Subcorpus 1	179	51	0,00102
Subcorpus 2	284	78	0,00156
Subcorpus 3	529	182	0,00364
Subcorpus 4	637	247	0,00494

Table 2. *Expected values (rounded at integers) of V(EV) and V1 (EV1) and the value of productivity (P) for the four subcorpora, calculated on the basis of the ZM model at the equal value of N=50000*

As it is shown in Table 2, once productivity is calculated at equal values of N, the results meet the expectation that the productivity of the prefixoid *samo-* in Russian increases over time. The vocabulary growth curves for each subcorpus at the token value of N=50000 are shown in Fig. 1:

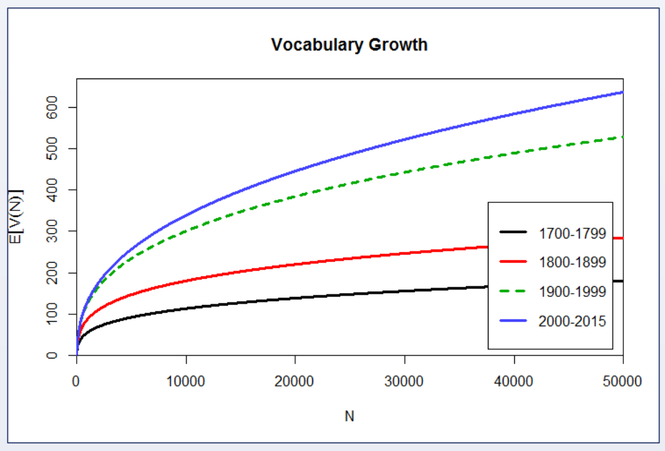

Fig. 1. *Vocabulary growth curves for the four subcorpora estimated on the basis of the ZM model at N=50000*

For each vocabulary growth curve, it is then possible to observe the confidence intervals (set by default to 95%), which allow to consider the variance estimates for the vocabulary size. Fig. 2 shows the confidence intervals around the vocabulary growth curves for each subcorpus at the token value of N=50000:

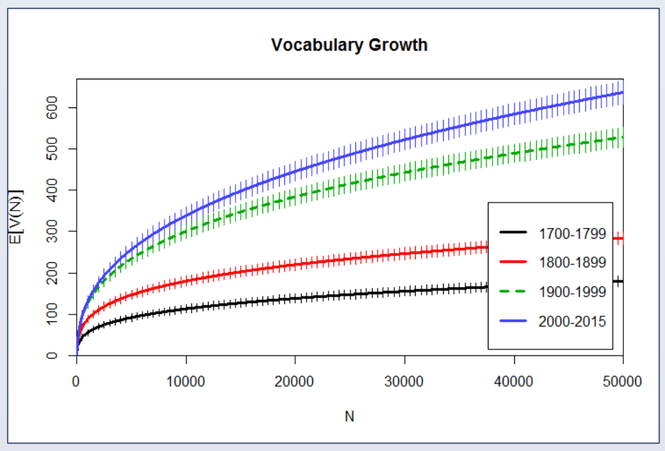

Fig. 2. *Confidence intervals at N=50000*

CONCLUSION

By adopting a corpus-based quantitative approach to the study of morphological productivity, I have provided empirical evidence of the increasing profitability of the prefixoid *samo-* in Russian in diachrony. The initial obstacle of comparing different-sized corpora has been overcome by employing parametric statistical models of frequency distribution (LNRE), which allow to estimate the number of types and *hapax legomena* with a given affix at arbitrary token values. The counterintuitive results obtained by applying the measure to calculate potential productivity to the collected data have thus been refuted because they are highly dependent on the corpus size. More trustworthy results have been obtained through the implementation of LNRE models. The prefixoid *samo-* in Russian shows an increasing productivity over time: this affix is still very active in Russian and it is currently employed to coin new words, many of which are not included in dictionaries yet.

REFERENCES

Baayen, H. R. (1992), *Quantitative aspects of morphological productivity*, in *Yearbook of Morphology 1991*, Booij, G. and van Marle, J. (eds.), Amsterdam: Springer, pp. 109-149.

Baayen, H. R. (1993), *On frequency, transparency and productivity*, in *Yearbook of Morphology 1992*, Booij, G. and van Marle, J. (eds.), Dordrecht: Kluwer, pp. 181-208.

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

Baayen, H. R. (2008), *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*, New York: Cambridge University Press.

Baayen, H. R. and Lieber, R. (1991), *Productivity and English derivation: a corpus-based study*, "Linguistics" 29-5, pp. 801-843.

Baroni, M. and Evert, S. (2014), *The zipfR package for lexical statistics: A tutorial introduction*, < http://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf>.

Plag, I. (2006), *Productivity*, in *The Handbook of English Linguistics*, Aarts, B. and McMahon, A. (eds.), Oxford: Blackwell, pp. 537-556.

Russian National Corpus, < http://ruscorpora.ru/>.