Morphological Productivity in Diachrony: The Case of the Deverbal Nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th Century

Pavel Štichauer
Charles University in Prague

1. Introduction

The present study[1] investigates, in a diachronic perspective, the productivity of three Italian deverbal suffixes within the framework of the corpus-based quantitative approach to morphological productivity (e.g., Baayen 1992, 2001, 2008). This approach conceives of productivity as the likelihood of observing a new formation with a given affix when sampling a large corpus. The main aim of the study is to support with empirical evidence the well-known claim, as formulated, for example, by Frenguelli (2005: 194), that there is a partial reduction of the productivity of *-mento*, the increasing profitability of the suffix *-zione* and the disappearance of *-gione* in Old Italian within the time span that goes from the 13th to the 16th century.[2]

To carry out the study, we have used the corpus LIZ 4.0 (*Letteratura Italiana Zanichelli*). LIZ 4.0 is a historical diachronic corpus (cf. Claridge 2008) which contains 1000 texts of differents genres from the 13th to the beginning of the 20th century; its overall size is roughly 40 millions of tokens, but each period is, of course, represented rather unequally. This is due mainly to the fact that for some periods (especially the early ones) some textual types are simply missing. The choice of this (rather unbalanced) corpus has been dictated by the need to cover the four periods from the 13th to the 16th century. In fact, this time span is not available for a larger corpus of Old Italian *ItalNet* (http://www.nd.edu/~italnet/; roughly 18 millions of words) which contains documents only up to the end of the 14th century. Moreover, LIZ 4.0 enables further division to subcorpora corresponding to centuries[3]. Thus, we have selected four subcorpora which represent the four time frames in question, each of the following sizes (cf. Table 1).

Table 1. *The size of the four selected subcorpora*

SUBCORPUS	CORPUS SIZE (in tokens)
13th century	732114
14th century	3565818
15th century	2675016
16th century	10604452

The main idea underlying this research is that the comparison between these different-sized corpora can be interpreted diachronically, the four corpora being of four different periods, much in the vein of

[1] I wish to thank Marco Baroni, Ingo Plag and Davide Ricca for their important remarks on the poster version of this paper. I am also grateful to two anonymous reviewers for their comments and observations to a previous version of the paper. This work is based on my monograph *La produttività morfologica in diacronia: i suffissi* -mento, -zione *e* -gione *in italiano antico dal Duecento al Cinquecento*, currently in press at Charles University Press, Prague [Karolinum, Praha], to appear in 2009. The work has been carried out within the research project n. 405/06/P009 supported by the *Czech Science Foundation* [Grantová agentura České republiky, GAČR].

[2] "...variazioni diacroniche che coinvolgono alcuni tipi di suffissati dalle Origini alla fine del XV secolo (per es. la progressiva scomparsa di N-*agione/-igione*; la riduzione di N-*anza/-enza*, (...), N-*mento*; l'incremento di N-*zione*..."

[3] Of course, this diachronic division is quite arbitrary, driven by the traditional periodization, and a further – narrower – subdivision could be realized. But this is left for future research.

© 2009 Pavel Štichauer. Selected Proceedings of the 6th Décembrettes, ed. Fabio Montermini, Gilles Boyé, and Jesse Tseng, 138-147. Somerville, MA: Cascadilla Proceedings Project.

Lüdeling & Evert (2005), who formulate the following "instructions": "First, determine the *synchronic* productivity of the process at a given point in time, using a statistical model that takes the stochastic nature of (synchronic) vocabulary growth into account. (...). Second, study the *diachronic* aspect of productivity by comparing the degree of synchronic productivity at two (or more) different points in time."

This paper is organized as follows. In section 2, we introduce the basic notions of the quantitative approach to morphological productivity. In section 3, we present the empirical data sampled from the four subcorpora in question. In section 4, we turn to a very important question, namely the problem of lemmatisation and elimination of types which are supposed not to pertain to the word-formation rule in question. In section 5, we present the results and their interpretation. Section 6 brings a brief conclusion devoted to specific problems connected with this approach.

2. The quantitative approach to morphological productivity

In order to adopt the corpus-based quantitative approach to productivity, it is important to introduce the well-known terminology as well as some basic notions and methods.

2.1. Terminology

The first important distinction is that of *types* and *tokens* (e.g. Baroni 2009). The words in a corpus can in fact be considered in two different ways. First, we can count the number of all words which are different with respect to what lexemes they represent. Thus the words like *am, is, are* represent the same lexeme *to be* and can therefore be subsumed under one single type. At the same time, the inflectional forms of the type *to be* can also be counted as instances of the given type and, in this sense, we speak about *tokens* of that type. Or, as far as complex words are concerned – as it is the case in the present study – we can count all different words with a given affix: for example, *acceptable, comparable, controllable, demountable, cleanable* are all words with the suffix *-able*. These five words represent five different types; each of them can occur in a corpus more than once and so the overall number of tokens can be much higher. As a matter of fact, only *cleanable* occurs in the *British National Corpus* precisely once while the others occur much more frequently. The overall number of tokens of these five types is, within the BNC, 5603.

The number of types is usually referred to as *type frequency* (noted henceforth as V) and is considered as an indicator of *vocabulary* (or *lexical*) *richness* of a given text or corpus. The number of tokens, on the other hand, is labeled *token frequency* (noted henceforth as N). The token frequency is important for two reasons. First, it is normally the overall number of tokens which is given as the corpus size (noted henceforth as F, following Baayen 1992). Second, the difference in token frequency can also play a significant role when comparing different type frequencies. In fact, some groups of formations may display a more or less identical type frequency, but they can be totally different with respect to their token frequency. In this sense, it may be important to use also *relative token frequencies* (noted henceforth as R_f), especially when we wish to carry out such a comparison across different-sized corpora (which is the case of this study). The relative frequency is calculated by N / F * 100 (or 1000), i.e. the number of tokens of the affix (N) divided by the overall number of tokens of the corpus (F) and multiplied with 100 (%) or 1000 (‰).

2.2. Notions

The relation of type frequency and token frequency is important in that the former (V) can be viewed as a function of the latter (N): the increasing value of N (given by the corpus size) will lead to the increasing value of V (cf. Baayen 1992: 113). As Baayen himself (2008: 222) puts it: "If we read through a text or corpus, and at regular intervals keep note of how many different types we have encountered, we find that, unsurprisingly, the number of types increases, first rapidly, and then more and more slowly." This relation gives rise to the notion of *vocabulary growth curve*.

Moreover, the rate at which the vocabulary grows can be captured by the proportion of *hapax legomena* (noted henceforth as V_1), the types that occur precisely once, to the overall number of tokens of the formations with a given affix, i.e. $P = V_1 / N$. This *vocabulary growth rate* is conceived of as the likelihood of observing a new type when sampling a corpus: "The growth rate is a probability, the probability that, after having read N tokens, the next token sampled represents an unseen type, a word that did not occur among the preceding N tokens." (Baayen 2008: 223).

2.3. Methodology

It is clear that the measure P, based on the relation V_1 / N, depends directly on the corpus size. It is therefore not possible to compare corpora of different sizes using this measure (cf. Baayen 1992: 117). In order to overcome this problem, two main techniques, which permit the unification of the value of N, are available. Either we can use *binomial interpolation*, or we can recur to one of the available models of *extrapolation*.

Binomial interpolation (cf. Baayen 2001: chap. 2) produces the *expected values* of V, V_1 (and other frequency classes), noted as E(V), E(V_1), for arbitrary values of N equal or smaller than the empirical value of N given by the corpus size (cf. Baroni & Evert 2006: section 3.3). The technique of extrapolation produces the *expected values* of V, V_1 for arbitrary values of N *larger* than the empirical value of N. The extrapolation relies on parametric statistical models of frequency distribution known as LNRE models (*Large Number of Rare Events*). Currently, three major models are available: *Generalized Inverse Gauss-Poisson* (GIGP; Baayen, 2001), *finite Zipf-Mandelbrot* and *Zipf-Mandelbrot* (fZM and ZM; Evert 2004; Baroni & Evert 2006). These models are implemented in the package *zipfR*, a tool for lexical statistics in R. The *zipfR* package, which will be used throughout this study, is being developed by Marco Baroni and Stefan Evert (Baroni & Evert 2006; http://zipfr.r-forge.r-project.org/)

3. The empirical data

We may now turn to the empirical data sampled from the four subcorpora. As we alredy mentioned, the types we are interested in are formed by three deverbal suffixes *-mento* (e.g. *cominciamento*, 'beginning'), *-zione* (e.g. *comparazione*, 'comparison') and *-gione* (e.g. *cacciagione*, 'hunt'). The relevant figures are summarised in the following table 2, where the number of types (V), the number of tokens (N) and the number of *hapax legomena* (V_1) are given. The relative token frequency (R_f), as defined above, is also presented.

Table 2. *The number of types (V), tokens (N) and hapax legomena (V_1) of the Italian suffixes* -mento, -zione *and* -gione *according to the given corpus*

13th century, corpus size (F) = 732114 tokens				
SUFFIX	**V**	**N**	**V_1**	**R_f (‰)**
-mento	280	2093	137	2.85
-zione	143	1653	47	2.25
-gione	22	116	9	0.15
14th century, corpus size (F) = 3565818 tokens				
SUFFIX	**V**	**N**	**V_1**	**R_f (‰)**
-mento	455	6475	206	1.81
-zione	398	7717	119	2.16
-gione	67	269	34	0.07
15th century, corpus size (F) = 2675016 tokens				
SUFFIX	**V**	**N**	**V_1**	**R_f (‰)**
-mento	351	3457	172	1.29
-zione	548	6679	177	2.49
-gione	24	62	12	0.02

16th century, corpus size (F) = 10604452 tokens				
SUFFIX	**V**	**N**	$\mathbf{V_1}$	$\mathbf{R_f}$ (‰)
-mento	583	14026	281	1.32
-zione	722	32031	194	3.02
-gione	31	233	13	0.02

Let us make some general remarks. The suffix *-mento* shows up, across the corpora, as a relatively constant affix: the number of types (V) increases with the increasing size of the corpus. Also, the proportion of *hapax legomena* to the overall number of types is constant. The suffix *-zione*, on the other hand, displays a diachronic variability in all reported values. Finally, the suffix *-gione* presents the values that turn out to be *incomparable*, and so *-gione* will be left out of the comparison (along the lines proposed by Gaeta & Ricca 2006: 84-85 who claim that one should give up comparing affixes with "extremely divergent" token frequencies.[4]

4. Principles of lemmatisation and type elimination

The empirical values that we have just presented rely on two important steps which can be considered as *linguistic prerequisites* to the calculation of the productivity (cf. Dal *et al.* 2007) and which turn out to be particularly tricky when dealing with a historical corpus (cf. Claridge 2008: 246-249): 1) Lemmatisation; 2) Elimination of lexemes that do not pertain to the rule in question.

1) As far as lemmatisation is concerned, we have decided to proceed *manually* so as to minimise, among other things, the *extraction noise* (cf. Evert 2005: 63). Although the corpus LIZ 4.0 is lemmatised, it is not at all reliable. Thus the only preprocessing consisted in a simple character-sequence query (e.g. .**mento/menti*, .**ione/ioni/ion*, etc.) which yielded a list of all the "lexeme-candidates". These forms have been checked in their contexts and lemmatised with special regard to variants which can be of two major types.
   The first type is rather unproblematic. It follows from the nature of the corpus that there may be a large number of *ortographic* variants. We have decided to fuse them wherever there was no doubt about the *type identity* of the form in question. By way of example, we can give the couple *benedictione* / *benedizione* which has been clearly subsumed under one single type BENEDIZIONE.
   The second type of variants is more problematic because it involves, in some cases, regional variation (for the discussion of this kind of problem cf. Claridge 2008: 249). Morphophonological variants such as the following ones *melioramento* / *miglioramento* / *meglioramento* ('amelioration, betterment') can in fact be considered as regional variants insofar as one assumes the closing of pretonic vowels (***mi**glioramento* instead of ***me**glioramento*) to be one of the typical features of the Florentine dialect of the 13th and the 14th century. But the question of the *type identity* of these different forms can be viewed as rather uncontroversial. In fact, we have decided to lemmatise all such morphophonological variants within one single type simply because otherwise there would have been a very large number of (sometimes) low frequency items which would have obviously distorted the results we wished to obtain.
2) As for the problem of what counts as an admissibile type which is to be included in the frequency counts, we have followed the same criteria already used for synchronic studies of Italian by Gaeta & Ricca (2002; 2003; 2006). The types that we have excluded from the frequency counts can be sorted in five groups.
   First, there are formations which display a strong semantic opacity with respect to a potential (or original) compositional meaning (e.g. *nazione*, 'nation'). Second, there are *baseless*

[4] The problem is that such an extremely divergent token frequency does not allow any statistically reliable "unification" of these values for mutual comparison, cf. below 5.1. For a thorough discussion of the suffix *-gione*, cfr. Štichauer (in press: chap. 4).

formations which can be of two types: either these formations are typical *mots complexes non construits* (in the sense of Corbin 1987), e.g. *elemento*, ‘element’, *venazione* ‘hunt’, or they belong to a small group of suppletive formations, e.g. *combustione*, ‘combustion’. Third, there is a rather numerous group of types whose base is not verbal; these formations have also a different meaning (collective, in most cases, e.g. *vasellamento*, ‘pottery’, *fogliamento*, ‘foliage’). Fourth, there is the group of formations which display derivational inner cycles. In line with what Plag (1999: 28-29) proposes, we have decided to exclude this kind of formations wherever a type in *-mento* or *-zione* constituted a base for a further complex word, e.g. *inconsiderazione*, ‘inconsideration’, *scorrezione*, ‘wrongness’. Finally, there is a small group of loanwords which have been excluded also wherever the verbal base was completely absent, e.g. *saramento*, ‘oath’, modelled on the French *serment*. The five groups are summarised in the following table 3 where some other examples are also given[5].

Table 3. *Examples of excluded types according to five criteria*

Type group - criteria	***Examples of excluded types***
1) Strong *opacity* (opaque semantic relation and/or absence of transpositional meaning)	*nazione* [‘nation’], *fazione* [‘faction’], *legazione* [‘legation’], *dormizione* [‘dormition’; *dormitio sanctae Mariae*], *sacramento* [‘sacrament’], *etc.*
2) *Baseless* formations (“mot complexe non construit” à la Corbin 1987, strong suppletivism)	*elemento* [‘element’], *argomento* [‘argument’], *tormento* [‘torment’], *atramento* [‘ink’], *venazione* [‘hunt’], *etc.*
3) Nominal base (or inexistant verbal base; different semantic instruction associated with the suffix)	*casamento* [‘house, building’], *vasellamento* [‘pottery’], *fogliamento* [‘foliage’], *drappamento* [‘clothes’], *palamento* [‘oarage’], *etc.*
4) Derivational inner cycles	*incorruzione* [‘uncorruption’], *inconsiderazione* [‘inconsideration’], *scorrezione* [‘wrongness’], *indeterminazione* [‘indetermination’] *etc.*
5) Borrowings	*saramento* (Fr. *serment*) [‘oath’], *dighisamento* (Old Fr. *desguisement*) [‘disguise’, ‘costume’], *etc.*

5. Comparison of the corpora

We are now ready to proceed to the comparison of the four corpora. First, we will mention the choice of one LNRE model that will be used throughout this section. Second, we will present the vocabulary growth curves at two unified values of N. Finally, we will also present the P values based on the expected values estimated from the four samples.

5.1. ZM model

The corpora being of different – *gradually increasing* – sizes, the technique of *extrapolation* has been used along the following lines.

The suffix *-gione* is left out of the comparison because the extrapolation from the values of N of this suffix to the values of the other two suffixes would exceed the reliable limit of extrapolation (for the discussion of the limits of reliable extrapolation, cf. Evert & Baroni, 2006).

In order to provide a unified coherent picture, only one LNRE model, implemented in the package *zipfR*, has been used throughout the study. The choice is also justified for two other reasons. First, on the basis of the experiments[6] analogous to the ones presented by Evert & Baroni (2006), the *Zipf-Mandelbrot* (ZM) model has been selected and the parameter estimation has been set to the *exact calculations* option because of the better statistical tests results (*goodness-of-fit, df, p-value*). Second,

[5] For a complete list of excluded types according to each period / corpus and to each suffix, cfr. Štichauer (in press: chap. 4).

[6] For the results of these experiments, cf. Štichauer (in press: chap. 4.2.3).

the ZM model performs much better even with small sample sizes, while the fZM and GIGP models tend to fail at parameter estimation (cf. Evert & Baroni 2006).

5.2. Vocabulary growth curves

As we have already seen, the range of sample sizes being too wide (from N = 1653 to N = 32031) for a reliable comparison, we are going to present the results at two unified values of N for all four corpora / periods. The first N = 7000 (in fig. 1.1., 1.2.) can certainly be considered as an acceptable value, while that of N = 20000 (in fig. 2.1., 2.2.) is to be taken as very approximate (the smallest empirical value of N is highlighted with the dashed vertical line).

Fig. 1.1. *Vocabulary growth curves for the suffix -* mento *in the 13th, 14th, 15th and the 16th century estimated within the* ZM *model at N = 7000*

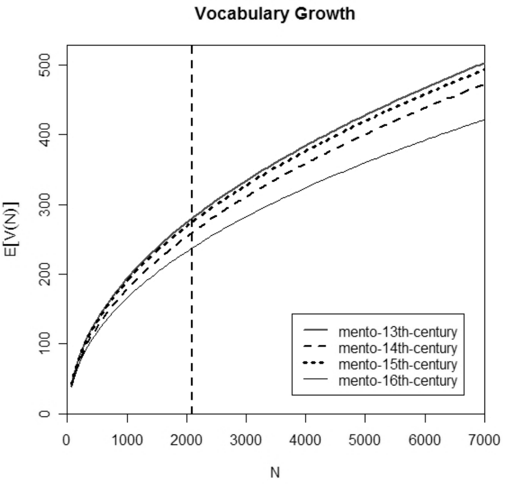


Fig. 1.2. *Vocabulary growth curves for the suffix -* zione *in the 13th, 14th, 15th and the 16th century estimated within the* ZM *model at N = 7000*

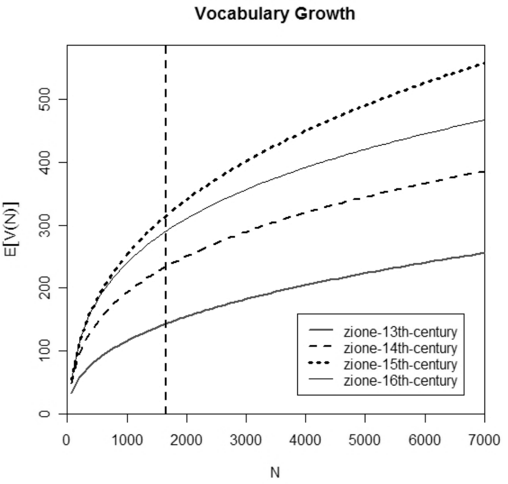


Fig. 2.1. *Vocabulary growth curves for the suffix -* mento *in the 13th, 14th, 15th and the 16th century estimated within the* ZM *model at N = 20000*

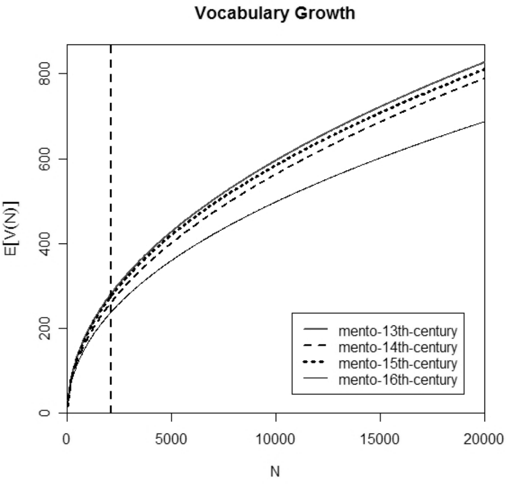
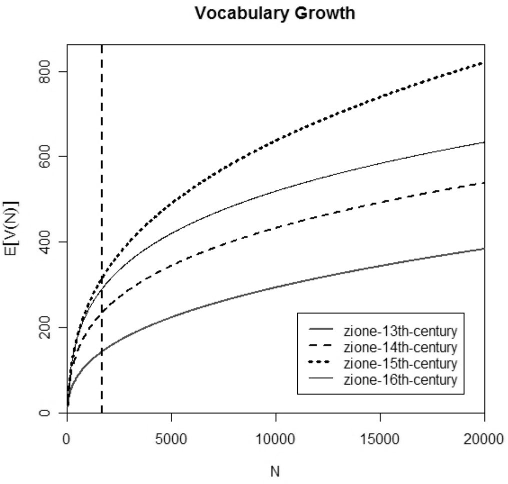


Fig. 2.2. *Vocabulary growth curves for the suffix -* zione *in the 13th, 14th, 15th and the 16th century estimated within the* ZM *model at N = 20000*



5.3. The values of P estimated on the basis of the ZM model

Although the visual comparison of the growth curves we have just presented allows to see some immediate results, we are also going to give the exact figures estimated on the basis of the ZM model. In the following two tables (Table 4., 5.) the *expected values* of V, V_1 are given as well as P values, calculated from $E(V_1)$ / N.

Table 4. *The expected values of V, V_1 (rounded to integers) and the value of P for the suffixes* -mento *and* -zione *from the 13th to the 16th century, estimated on the basis of ZM model at the equal value of N = 7000*

SUFFIX	**N**	**E(V)**	**E(V_1)**	**P (V_1/N)**
-mento (13th century)	7000	503	240	**0.0342**
-mento (14th century)	7000	473	232	**0.0331**
-mento (15th century)	7000	493	236	**0.0337**
-mento (16th century)	7000	421	198	**0.0282**
-zione (13th century)	7000	255	100	**0.0142**
-zione (14th century)	7000	386	126	**0.0180**
-zione (15th century)	7000	558	211	**0.0301**
-zione (16th century)	7000	467	142	**0.0202**

Table 5. *The expected values of V, V_1 (rounded to integers) and the value of P for the suffixes* -mento *and* -zione *from the 13th to the 16th century, estimated on the basis of ZM model at the equal value of N = 20000*

SUFFIX	**N**	**E(V)**	**E(V_1)**	**P (V_1/N)**
-mento (13th century)	20000	827	390	**0.0195**
-mento (14th century)	20000	789	383	**0.0191**
-mento (15th century)	20000	811	382	**0.0191**
-mento (16th century)	20000	687	318	**0.0159**
-zione (13th century)	20000	384	147	**0.0073**
-zione (14th century)	20000	538	167	**0.0083**
-zione (15th century)	20000	821	295	**0.0147**
-zione (16th century)	20000	634	178	**0.0089**

5.4. What the results show

We believe that the major, though modest, outcome of this research is simply that it brings a new kind of evidence to the well-known claims mentioned above. Indeed, from the visual comparison of the vocabulary growth curves, we can draw the following conclusions. The suffix -*mento* shows up, in diachrony, as productive in a rather constant way, only in the 16th century does it show, effectively, some light decrease, in line with the above mentioned claim. The suffix -*zione*, on the other hand, reaches its productivity peak in the 15th century, which is the period of strong influence of Latin, and still scores high in the 16th century. Further estimation of the values of *V*, *V_1* and the *P* value offered by the ZM model confirms the more or less constant vocabulary growth rate for the suffix -*mento* and the variable rate for the suffix -*zione*.

Moreover, the fact that the suffix -*mento* can be considered as diachronically constant (with the exception of the 16th century) and that the suffix -*zione*, on the other hand, shows up as diachronically variable can also be seen when we compare the *confidence intervals* around the vocabulary growth curves[7] in figures 3.1. and 3.2.

[7] I wish to thank Ingo Plag for having pointed out this to me. The confidence level is set by default to 95%.

Fig. 3.1. *Vocabulary growth curves for the suffix -* mento *in the 13th, 14th, 15th and the 16th century estimated within the* ZM *model at N = 20000 with the indication of the 95% confidence intervals*

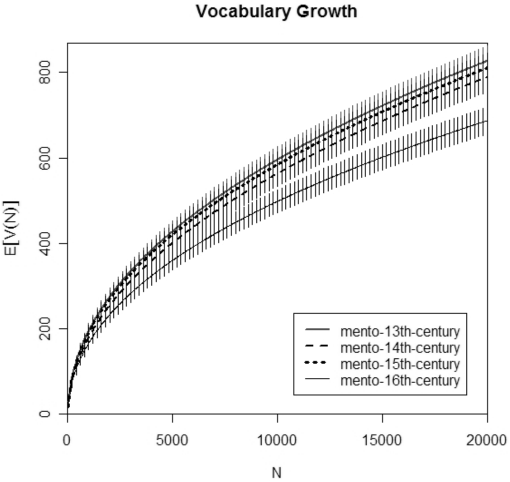
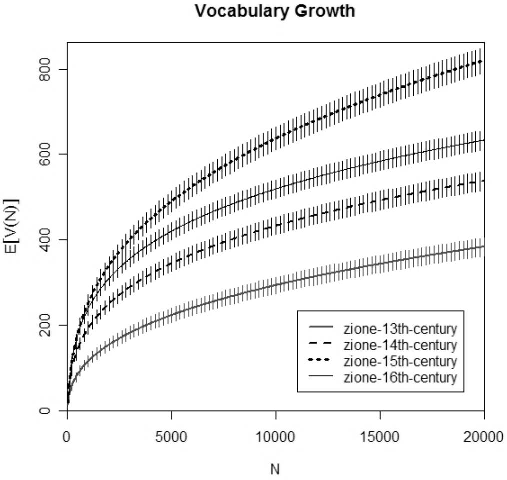


Fig. 3.2. *Vocabulary growth curves for the suffix -* zione *in the 13th, 14th, 15th and the 16th century estimated within the* ZM *model at N = 20000 with the indication of the 95% confidence intervals*



In fact, the overlap between the confidence intervals of the suffix *-mento* in the 13th, 14th and the 15th century suggests that there is no significant difference between these curves and that the suffix *-mento* is, across these three centuries, essentially identical in its productivity. As far as the suffix *-zione* is concerned, the fact the the confidence intervals are outside each other shows clearly that there is significant diachronic variation.

However, this purely quantitative estimation of the productivity of these two rival suffixes does not permit to see any particular aspects of the presumed rivalry. If there is a strong correlation between the light decrease of the productivity of *-mento* in the 16th century and a substantial increase of *-zione*, we still do not see any precise cases where this rivalry reveals itself. What we need therefore is a detailed analysis of cases where the same verbal base displays a diachronically different selection of the suffix. In fact, there are such cases, e.g. *consolamento* / *consolazione* ‘consolation’, where we see that the former is gradually substituted with the latter form. It is also in this context of competition that the quantitative measure we have been working with may become interesting in a complementary way (cf. Lüdeling *et al.* 2006).

6. Conclusions

To conclude, I wish to mention some problems typical of this approach in general but which can be seen as even more important in the case of our diachronic research. There are at least two problems with this approach when applied to diachronic corpora, one regarding the process of lemmatisation and elimination, the other the strong non-randomness of a diachronic corpus.

The process of lemmatisation / elimination turns out be tricky for two reasons. First, some formations – typically those in *-zione* – appear in the corpus without the verbal base which enters the Italian language a little later, e.g. *asserzione – asserire*, ‘assertion – assert’, or *inquisizione – inquisire*, ‘inquisition – inquire’. In Old Italian, *inquisizione* is found in couple with the latinate verb *inquirere*. Only later on is the verb *inquisire* derived by the process of back-formation. For this particular reason, these formations, although clear borrowings from Latin, are included in our lists of types. Second, some formations could have been “fused” into one type, e.g. *disprezzamento – dispregiamento*, ‘despise, contempt’, *soddisfacimento – satisfacimento*, ‘satisfaction’. In the present study, though, these formations have been kept as separate types whenever the corresponding verb was also present.

The non-randomness is due to various factors, two of them being highly typical of the diachronic corpus: 1) inhomogeneity and 2) clustering / repetition effects (cfr. Evert 2005: 59; 2006).

1) Inhomogeneity. The diachronic corpus on which this study has been conducted is made up from documents with different properties (different textual typology, different authors, and different proportions of the text types across the four subcorpora, which is due to the fact that some text types, in a given period, are simply missing). It may be therefore objected that diachronic corpora such as this one are completely unusable because they cannot reach any "reasonable degree of statistical validity" (Claridge 2008: 247).
2) Clustering / repetition effects. This kind of problem is more serious. It is mainly due to the fact that most of the texts are author's texts and so the distribution of certain formations tends to display strong repetition effects in this sense. A typical – and, unfortunately, frequent – case is when an author coins a new formation – a hapax legomenon candidate – that he uses more than once throughout the text; or when a high-frequency item is not distributed evenly across the corpus but is concentrated in one or two concrete documents.[8]

There are some solutions / tests of variation of different sorts (the comparison of empirical distribution of frequencies with theoretical ones, the dispersion test, etc., cfr. Evert 2006: section 5), but this kind of assessment and validation requires further work[9].

References

Baayen, Harald. 1991. Quantitative aspects of morphological productivity. In G. Booij and J. van Marle, eds. *Yearbook of Morphology 1991*. Dordrecht: Kluwer. 109-149.

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.

Baroni, Marco. 2009. Distributions in text, in A. Lüdeling and M. Kytö, eds. *Corpus Linguistics. An International Handbook*, vol. 2, article 37. Berlin: Mouton de Gruyter. 803-822.

Baroni, Marco and Stefan Evert. 2006. The *zipfR* package for lexical statistics: A tutorial introduction (http://www.cogsci.uni-osnabrueck.de/~severt/zipfR/).

Claridge, Claudia. 2008. Historical Corpora. In A. Lüdeling and M. Kytö, eds. *Corpus Linguistics. An International Handbook*, vol. 1, article 14, Berlin: Mouton de Gruyter. 242-259.

Corbin, Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique*, 2 voll. Tübingen: Niemeyer.

Dal, Georgette, Bernard Fradin, Natalia Grabar, Stéphanie Lignon, Fiammetta Namer, Clément Plancq, François Yvon and Pierre Zweigenbaum. 2007. Linguistic prerequisites to the calculation of morphological productivity and first results. Paper delivered at *Journées ATALA*, Paris, November 10, 2007.

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; available from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/.

Evert, Stefan. 2006. How Random is a Corpus? The Library Metaphor. *Zeitshrift für Anglistik und Amerikanistik*, 54.2: 177-190.

Evert, Stefan and Marco Baroni. 2006. Testing the extrapolation quality of word frequency models. In *Proceedings of Corpus Linguistics 2005*. www.corpus.bham.ac.uk/PCLC/.

[8] For example, *consolazione* [consolation] has 246 tokens in the corpus of the 14th century of which 60 tokens are found in Caterina da Siena's *Lettere*. A more "drastic" proportion is found with *raccomandazione* [recommendation] that occurs in the 16th century 144 times, but as many as 138 tokens are found in Torquato Tasso's *Lettere*.

[9] Actually, the comparison of empirical *frequency spectra* with those obtained by ZM model has been carried out. The results differ according to each sample the worst result being for the suffix *-zione* in the 15th century which reveals, undoubtedly, some problems with the underlying corpus.

Frenguelli, Gianluca. 2005. Nominalizzazione e testualità nella trattatistica del XV secolo. In M. Grossmann and A.M. Thornton, eds. *La formazione delle parole. Atti del XXVII Congresso Internazionale di Studi della Società di Linguistica Italiana. L'Aquila, 25-27 settembre 2003*. Roma: Bulzoni. 193-209.

Gaeta, Livio and Davide Ricca. 2002. Corpora testuali e produttività morfologica: i nomi d'azione in due annate della *Stampa*. In R. Bauer and H. Goebl, eds. *Parallela IX. Testo – variazione – informatica. Text – Variation – Informatik*. Wilhelmsfeld: Gottfried Egert Verlag. 223-249.

Gaeta, Livio and Davide Ricca. 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 Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44.1: 57-89.

LIZ 4.0. = Stoppelli, Pasquale and Eugenio Picchi, eds. 2001. *Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*. Bologna: Zanichelli.

Lüdeling, Anke and Stefan Evert. 2005. The emergence of non-medical *-itis*. Corpus evidence and qualitative analysis. In S. Kepser and M. Reis, eds. *Linguistic Evidence. Empirical, Theoretical, and Computational Perspectives*. Berlin: Mouton de Gruyter. 315-333.

Lüdeling, Anke, Marco Baroni and Stefan Evert. 2006. Need and competition: deconstructing quantitative productivity. Paper delivered at *Quantitative Investigations in Theoretical Linguistics 2 (QITL-2)*. University of Osnabrück, Germany, June 1-2, 2006.

Plag, Ingo. 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.

Štichauer, Pavel. In press. *La produttività morfologica in diacronia: i suffissi* -mento, -zione *e* -gione *in italiano antico dal Duecento al Cinquecento*. Praha: Karolinum.

Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux

edited by Fabio Montermini,
Gilles Boyé, and Jesse Tseng

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-433-1 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Štichauer, Pavel. 2009. Morphological Productivity in Diachrony: The Case of the Deverbal Nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th Century. In *Selected Proceedings of the 6th Décembrettes*, ed. Fabio Montermini, Gilles Boyé, and Jesse Tseng, 138-147. Somerville, MA: Cascadilla Proceedings Project.

or:

Štichauer, Pavel. 2009. Morphological Productivity in Diachrony: The Case of the Deverbal Nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th Century. In *Selected Proceedings of the 6th Décembrettes*, ed. Fabio Montermini, Gilles Boyé, and Jesse Tseng, 138-147. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2241.