International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

Comparing Lexicons Diachronically in Italian Literary Corpora

Luca Pavan
Institute of Foreign Languages, Vilnius University, Vilnius, Lithuania
Language Studies Centre, Faculty of Creative Industries, Vilnius Tech, Vilnius, Lithuania
Department of Foreign Languages, Literary and Translation Studies, Vytautas Magnus University, Kaunas, Lithuania
✉ **Corresponding Author**: Luca Pavan, **E-mail**: pavan@panservice.it

ARTICLE INFORMATION

Received: July 12, 2021
Accepted: August 21, 2021
Volume: 4
Issue: 8
DOI: 10.32996/ijllt.2021.4.8.13

KEYWORDS

Corpus linguistics, Diachronic corpora, Florentine vernacular literature, Italian literature

ABSTRACT

The goal of the article is to provide a comparison between several words from Florentine vernacular language and modern Italian language, using software written by the author. This paper focuses on two corpora: the first one includes a selection of Florentine vernacular literature and the second one a group of literary books written in a modern Italian language from the end of XIX Century up until the present. The article demonstrates the use of some features of the software to compare the two corpora, ranking the lexicographic entries using different strategies. It is possible to analyse the lexicon taking into consideration different types of sorting, using only three parameters: the word frequency, the percentage of frequency according to the number of words in the corpus, and the percentage of texts where the word is found in the corpus. From these parameters a fourth parameter also arises the level of persistence of words in each corpus. The software allows observing the differences in the use of lexicon in various periods of history, comparing the Florentine vernacular language, which was used in the Italian peninsula till the beginning of XIX Century, to the modern Italian language.

1. Introduction

The diachronic linguistics became, over the years, a promising field for corpus linguistics. Analysing and comparing corpora under a diachronic point of view can better understand language evolution over time. The researcher often moves from an overall word frequency analysis to a closer textual reading (Alessi, Partington, 2020, p. 9), a method that also involves statistical analysis.

The case of the Italian language can be studied with the aim of corpus linguistics. The language of literature, in fact, has a strong connection with Florentine vernacular language, which was used at least from the XIV Century till the beginning of XIX Century. The structure of modern Italian is similar to Florentine vernacular, although there are some differences. The 1827 edition of the work *I promessi sposi* by Alessandro Manzoni made use of a new language closer to modern Italian. Manzoni, already far from Florentine vernacular, intended to adopt the language for cultured readers and all others capable of reading (Dotti, 2020, p. 373). The lexicon also changed along history, but the majority of words coming from Florentine vernacular are still in use today, in the modern Italian language. Other words became obsolete and disappeared from modern Italian. A number of corpora built using modern Italian are available today (Rossini Favretti, 2002, p. 28), but corpus-based studies comparing the literature written in modern Italian and the literature written in Florentine vernacular are still missing. An attempt to compare Italian language corpora with texts from different epochs was already provided by the author (Pavan, 2020). The program CorpStat, a software written by the author, is used to conduct a corpus-based analysis. However, in this article, software packages like AntConc (Anthony, 2014) used to retrieve keywords will not be considered.

2. Method

Some features of CorpStat were already described in a previous article (Pavan, 2020). The software was used to analyse two corpora, both assembled by the author. The first corpus is a selection of works written in Florentine vernacular language, from XIV to XVIII Century. In this corpus there are mainly works of literature, including about 2,700,000 words. The second corpus is a selection of literary works from the end of XIX Century till today, which includes about 2,500,000 words. The sizes of both corpora are quite similar: comparing diachronic corpora should involve the use of corpora with similar size (Kaunisto, p. 3). Some major works of the XIX Century (like Manzoni's *I promessi sposi*) were intentionally not included in this corpus, assuming at that time the modern Italian language was not yet well established. Both corpora were at first tokenized by CorpStat. Three parameters are showed by CorpStat after the tokenization: the word frequency in the corpus, the percentage of frequency according to the number of words in the corpus, and the percentage of texts where the word is found in the corpus. The words are later sorted in different ways according to each parameter. For example, if the parameter taken into consideration is the word frequency in the first corpus, CorpStat sorts like this:

Sorting of frequency in the first corpus in descending order.
Only the words found in both corpora are sorted.
Total amount of words in the dictionary: 151101
Total amount of words in all corpora: 5241744

The list in the format a[b][c] includes:
a) word frequency in the corpus
b) percentage of frequency according to the number of words in the corpus
c) percentage of texts where the word is found in the corpus

== Columns ==
Column 1: Rank
Column 2: Word
Column 3: Corpus 1 - Works written in Florentine vernacular language --> Total number of words in this corpus: 2721608
Column 4: Corpus 2 - Works written in modern Italian --> Total number of words in this corpus: 2520136

1	*e*	115886 [4.258000][100.000000]	71010 [2.817710][100.000000]
2	*che*	86949 [3.194770][100.000000]	61597 [2.444190][100.000000]
3	*di*	51729 [1.900680][100.000000]	80981 [3.213360][100.000000]
4	*la*	51051 [1.875770][100.000000]	53151 [2.109050][100.000000]
5	*il*	47513 [1.745770][97.500000]	51558 [2.045840][100.000000]

In another example, if the parameter taken into consideration is the percentage of texts of the first corpus, in which the word is found, the output of CorpStat would be the following:

Sorting the percentage of texts where the word is found in corpus 1 (descending order).
……..

1192	*veste*	164 [0.006030][62.50]	78 [0.003100][56.10]
1193	*vivendo*	78 [0.002870][62.50]	29 [0.001150][43.90]
1194	*vivere*	184 [0.006760][62.50]	387 [0.015360][95.12]
1195	*volo*	189 [0.006940][62.50]	216 [0.008570][85.37]
1196	*accesa*	128 [0.004700][60.00]	74 [0.002940][73.17]
1197	*accesi*	97 [0.003560][60.00]	44 [0.001750][43.90]

Another kind of sorting is given by the *fork* between two parameters (or their difference): for example, taking into consideration the parameter [c] in both corpora, CorpStat sorts like this (the forks become smaller in descending order and they are sorted according to the first corpus):

Sorting the forks of percentages of texts where the words is found in corpora [c] (descending order).
Only the words found in both corpora are sorted.
……..

1	*a'*	2005 [0.073670][**97.50**]	13 [0.000520][**2.44**]

2 *meco* 641 [0.023550][**92.50**] 6 [0.000240][**4.88**]
3 *perch'* 449 [0.016500][**92.50**] 2 [0.000080][**4.88**]
4 *avea* 4048 [0.148740][**87.50**] 5 [0.000200][**2.44**]
5 *gentil* 511 [0.018780][**87.50**] 1 [0.000040][**2.44**]

In the previous example, to show how the sorting of forks works, it is possible to check the differences between the two parameters [c]:

97.50-2.44=95.06
92.50-4.88=87.62
92.50-4.88=87.62
87.50-2.44=85.06
87.50-2.44=85.06

It is also useful to draw a chart with these values (Fig. 1). In this case, it is possible to compare the words visually to check their *level of persistence*, moving diachronically from a corpus to the other one. A high value in Y-axis means that the word is less popular (or absent) in the second corpus. Conversely, a low value shows that the word is found in both corpora to a certain degree. It is also possible to invert the first corpus with the second one, getting opposite results if it is more comfortable.

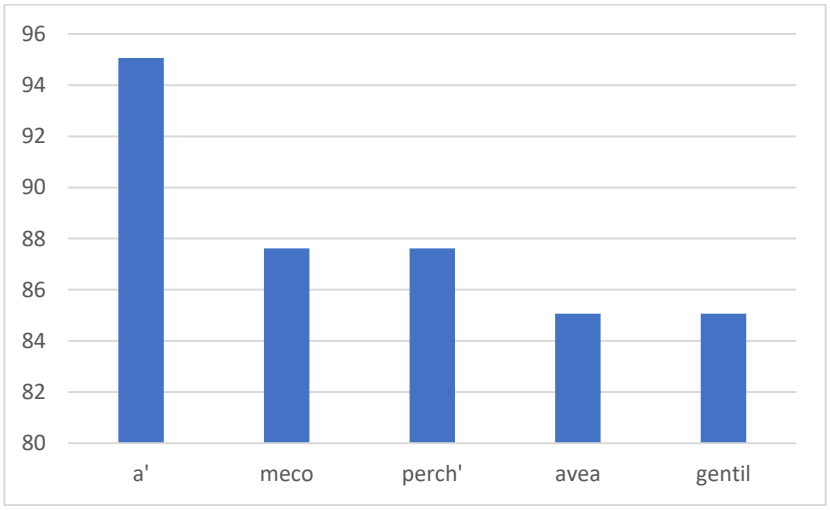

Fig. 1 – *Chart showing the level of persistence for a group of words*

It is possible to get forks for other parameters too. So, taking into consideration the parameter [b], the forks are sorted like this:

Sorting the forks of the percentages of words' frequencies in the corpora [b] (descending order).
Only the words found in both corpora are sorted.
……..

25 *avea* 4048 [**0.148740**][87.50] 5 [**0.000200**][2.44]
26 *voi* 4978 [**0.182910**][97.50] 941 [**0.037340**][97.56]
27 *ben* 4709 [**0.173020**][100.00] 808 [**0.032060**][100.00]
28 *ne* 8585 [**0.315440**][100.00] 4435 [**0.175980**][100.00]
29 *son* 3885 [**0.142750**][100.00] 120 [**0.004760**][43.90]

On the other hand, sorting the forks according to [b] in the second corpus gives an output like this:

1 *di* 51729 [**1.900680**][100.00] 80981 [**3.213360**][100.00]
2 *un* 15089 [**0.554410**][100.00] 37092 [**1.471830**][100.00]
3 *una* 8296 [**0.304820**][100.00] 24878 [**0.987170**][100.00]
4 *era* 7258 [**0.266680**][97.50] 22545 [**0.894590**][100.00]
5 *aveva* 1083 [**0.039790**][65.00] 11119 [**0.441210**][100.00]

3. Results and discussion

Collecting the data from CorpStat gives the opportunity to observe the differences of lexicons in both corpora of written language in the Italian peninsula along history. For example, it is possible to study the historical persistence of words belonging to the same grammatical gender like the pronouns, using the differences between parameters. During the era of Vernacular language several linguists wrote treatises about the grammar: Giacomo Pergamini, writing about pronouns, listed among them *questo*, *costui*, *colui*, *medesimo*, *esso* (Pergamini, 1626, p.79). Sorting these words with CorpStat, a list of forks between the parameters [c] would look like this:

186	*costui*	774 [0.028440][**77.50**]	35 [0.001390][**26.83**]
453	*colui*	1235 [0.045380][**90.00**]	113 [0.004480][**51.22**]
638	*esso*	1153 [0.042360][**87.50**]	114 [0.004520][**53.66**]
6726	*medesimo*	712 [0.026160][**55.00**]	44 [0.001750][**51.22**]
14527	*questo*	8018 [0.294610][**100.00**]	471 [0.177410][**100.00**]

While the differences of parameters [c] would look like this:

77.50-26.83=50.67
90.00-51.22=38.78
87.50-53.66=33.84
55.00-51.22=3.78
100.0-100.0=0.0

As shown earlier, it is possible to draw a chart (Fig. 2).

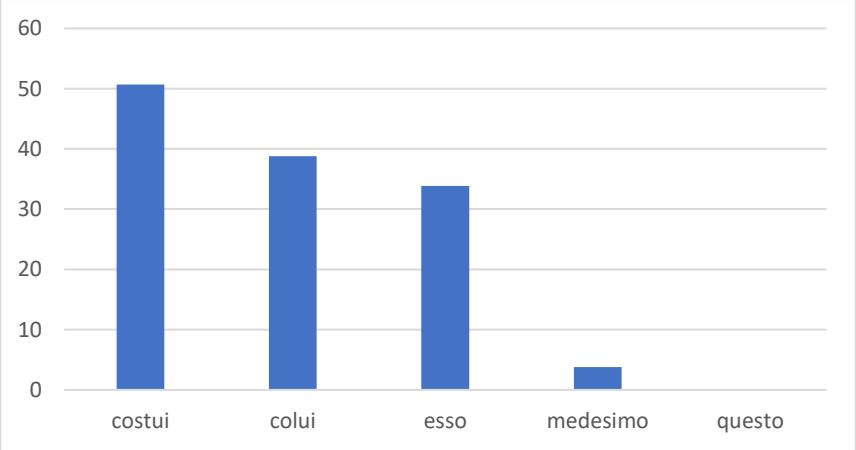

Fig. 2 – *Chart showing the level of persistence for a group of words*

Looking at the chart, the pronoun *costui* is less used in modern Italian when compared to the Florentine vernacular. *Colui* follows in the chart which means it was more popular in the Vernacular language. *At present, Esso is mostly avoided in writings and* is often replaced by *lui*, especially in the current Italian language. However, in the XIX century, *esso* was still used to some degree and this fact would explain its rank in the analysis. *Medesimo* and especially *questo* are very popular in both corpora. However, they can also have the function of being used as adjectives. For this reason, they also have more chances to be present in the corpora.

With the same method, many different kinds of analysis could be performed: for example. One could compare a group of adjectives, prepositions, nouns etc., to each other to make conclusions about the persistence of words along the timeline of history. Furthermore, CorpStat sorts the forks of parameters in both corpora, observing the ranking of words: in this case, one would want to get the most popular words in both corpora, or the less frequent words in one corpus.

In comparison with other languages, the modern Italian language has strong connections with his old ancestor, the Florentine vernacular, because of its similar language structure. For example, the old English looks more complicated in comparison with modern English, by the presence of unfamiliar words and spelling variants (Weisser, 2016, p. 15). In this case, CorpStat cannot compare words in both corpora since the software cannot detect the modifications of words and history. But the Italian language offers the opportunity to analyse the lexicon diachronically quite easily.

However, CorpStat is able to analyse more than two corpora, as it was already demonstrated in a previous article (Pavan, 2020). In this case the words' level of persistence could be easily drawn as lines between a couple of corpora. In addition, words and their modifications can be compared. This kind of analysis is quite common in corpus linguistics to check the frequency of old and new words (Jones, Waller, 2015, p. 30).

4. Conclusion

Corpus linguistics is important in lexicography to make, among other things, an inventory of a language's lexicon (Zufferey, 2020, p. 3). Software packages, like CorpStat, can build the lexicon at the same time showing the changes in language over time. The three parameters in the output of CorpStat can help define the words' modifications in the language and history. In fact, analysing the two corpora with the software described in the article, it is possible to compare the words diachronically. For the first time, the article introduces a special parameter - the level of persistence of words in a language, showing how much the words changed over centuries. However, the software described here has some limitations, especially if one wants to compare spelling evolution in a language other than Italian. In the future, the new versions of the software could include some capabilities for spelling, allowing analysing different languages. To understand the evolution of languages, the study of corpora needs instruments like the one described here to analyse the modifications in different historical periods**.**

References

[1] Alessi, G., Partington, A. (2020) *Modern diachronic corpus-assisted language studies: methodologies for tracking language change over recent time*. Fidenza (PR): Mattioli 1885.
[2] Anthony, L. (2014). AntConc 3.4.3, software. Tokyo: Waseda University.
[3] Dotti, U. (2020). *Storia della letteratura italiana*. Roma: Carocci Editore.
[4] Jones, K., Waller, D. (2015). *Corpus Linguistics for Grammar*. New York: Routledge.
[5] Kaunisto, M. (2007). *Variation and Change in the Lexicon*. Amsterdam: Rodopi.
[6] Pavan, L. (2020). Diachronic Analysis of Italian Opera Librettos. *International Journal of Linguistics, Literature and Translation*, *3*(4), 26-32. Retrieved from https://al-kindipublisher.com/index.php/ijllt/article/view/404
[7] Pergamini, G. (1626). *Trattato della lingu*a. Venezia: Ciotti.
[8] Rossini Favretti, R. (2002). *Corpus linguistics in Italian studies*, in Nuccorini, S. (ed.), Phrases and Phraseology - Data and Descriptions, Bern: Peter Lang.
[9] Weisser, M. (2016). *Practical Corpus Linguistics*. Malden (MA): Wiley & Sons.
[10] Zufferey, S. (2020). *Introduction to Corpus Linguistics*. London: Wiley-ISTE.