

Original Research Article

Diachronic Analysis of Italian Opera Librettos

Luca Pavan

Vilnius University, Faculty of Philology, Institute of Foreign Languages, Vilnius, Lithuania

Vilnius Gediminas Technical University, Faculty of Creative Industries, Language Studies Center, Vilnius, Lithuania

Corresponding Author: Luca Pavan, E-mail: pavan@panservice.it

ARTICLE INFO

Article History

Received: March 12, 2020

Accepted: April 11, 2020

Volume: 3

Issue: 4

DOI: 10.32996/ijllt.2020.3.4.3

KEYWORDS

Diachronic linguistics, sub-corpora, keywords, opera librettos

ABSTRACT

The purpose of this article is to find out which are the most typical words used in Italian opera librettos in each of historical periods of opera and to see the evolution of their use. A word which statistically occurs in a corpus of texts more often than expected is called a keyword. The article shows a new approach in diachronic linguistics, proposing a different examination of keywords found with the Log-Likelihood method in a corpus of Italian opera's librettos. To get the keywords, the libretto's corpus was compared with a reference corpus which includes the most representative works of Italian literature. Later the librettos' corpus was divided diachronically in 5 sub-corpora according to some historical periods. A new analysis of these sub-corpora was performed using a software tool written by the author called CorpStat. It allows to analyze the evolution in the use of lexicon of opera's librettos. Often the Log-Likelihood statistical method was criticized mainly because some words qualifies as keywords, even if they appear only in one or very few texts. With the software CorpStat it was possible to validate words as real keywords taking into consideration the percentage of number of texts in which they appear. The analysis brings to the conclusion that only 41 keywords out of a list of first 200 keywords could be validated as real keywords. However, the diachronic analysis shows that a word is a real keyword only in a certain historical period, but not necessarily in another. This can be due to the changes of libretto's plots and language along the history.

Introduction

The study of the language evolution through time is called diachronic or historical linguistics. It has developed especially in the last decades (Kytö, 2011, p. 417). The presence of historical corpora on Internet significantly simplified the research and the collection of historical texts. The language of the Italian opera's librettos can be studied comparing a certain number of librettos coming from different historical periods. There are also several opera librettos corpora to be found online.

To find out what are the most typical words used in opera librettos at first it is necessary to search for keywords and later to check whether they are also could be regarded as a keyword in a smaller number of librettos. For both steps the corpus of opera librettos is necessary. The author analyses the lexicon of a corpus consisting of about 300 Italian opera's librettos (www.librettidopera.it). This corpus can be defined as a *monogeneric corpus* (Partington, 2011, pp. 35-36) because includes a special kind of literature having particular features. Since the birth of opera, the librettos' language is strongly connected with poetry (Bonomi, Buroni, 2010, p. 9). In fact, until the half of XIX

century the language of the Italian opera's librettos derived from Florentine Vernacular language (Staffieri, 2014, p. 19-20) which was the language of poetry.

The search of keywords is done by using the statistical method which searches for words in libretto corpus occurring more often than statistically expected. Keywords, in a quantitative meaning, are those words whose frequency is statistically significant (Bondi, Scott, 2010, p. 3). To obtain the keywords it was necessary to compare two corpora: the librettos corpus mentioned above and a reference corpus, containing several works of the Italian literature. The latter corpus was composed by the author. After the discovery of the keywords the author selects the first 200 keywords. The goal is to validate these keywords according to their presence in a certain number of texts. To do this, the corpus of librettos was divided in 5 sub-corpora referencing the appropriate historical period. The threshold to validate the keyword is set by the author 20 % and more of selected keyword's presence in the texts of at least one sub-corpus. Finally, the 41 validated keywords are compared diachronically through these 5 sub-corpora.

The analysis shows that some librettos' words are more typical for a specific period rather than for another. The reason for this might be the change of librettos' plots used in the different periods.

Methodology

The keywords in the corpus of Italian opera's librettos were found by using the Log-Likelihood statistical method and with a help of the software AntConc (Anthony, 2014). This software tool, among other things, allows to compare a corpus of texts with another reference corpus to extract the keywords of the first corpus. The reference corpus, assembled by the author, includes many works mainly of Vernacular literature, starting from the XIV century. The software also ranks the keywords.

After the finding the keywords and after the selection of first 200 according to their rank, the corpus of librettos was divided in 5 sub-corpora, according to some historical periods. These sub-corpora were analyzed by using another software tool written by the author, called CorpStat, written in C programming language. This software tool is able to compare several corpora showing three parameters: the word frequency, the percentage of word frequency according to the amount of words in the corpus, and the percentage of the amount of texts in the corpus where the word appears at least once. These parameters are shown in this format:

a [b] [c]

a = word frequency

[b] = percentage of word frequency according to the amount of words in the corpus

[c] = percentage of the amount of texts in the corpus where the word appears

For example, if three corpora are compared, for each word the output will be like this:

abbandonata 21 [0.00261][30.50847] 51 [0.00813][26.92308] 28 [0.00370][22.85714]

The first column shows words sorted alphabetically. The other columns show the three parameters previously described for each corpus. The first parameter – word frequency - is a practical measure for a single corpus (Brezina, 2018, p. 42). The other two parameters, especially the third one, are useful to compare more corpora. CorpStat can also print results only in a given range of values. It is possible to assign a range of values to one parameter at a time. To achieve the results described in this article, CorpStat at first makes a token segmentation of the librettos' corpus. It is known that to decide which word is a token and which is not is never an easy task (Lüdeling, Kytö, 2009, pp. 804-805). In the case of CorpStat, after the punctuation is eliminated, the words are

converted in lower case and those words followed by an apostrophe are separated in two tokens. After that, the software performs the word counting in the corpus and the sub-corpora. Finally, the calculations of frequency and percentages are made in the output file.

Among the first 200 keywords coming from AntConc, some keywords were validated according to the parameter [c] coming from CorpStat. It was set a value of [c] = 20 %. Below this threshold the word is not selected. The word, to be chosen, must have [c] bigger than 20 % in at least one sub-corpora. Therefore, it is supposed that a word is typical only if it appears in at least the 20 % of texts in one or more sub-corpora. This minimum value of [c] is arbitrary and of course could be changed for a different kind of analysis.

The validated keywords, now called *real keywords*, constitute a new list of typical words in the librettos. However, comparing the sub-corpora with CorpStat, in some sub-corpora words have [c] less than 20 %. In this case the words are not real keywords in those sub-corpora. The list of real keywords contributes to build a lexicon used in the librettos taking into consideration its historical modifications.

The Making of Sub-corpora

For the diachronic analysis, the sub-corpora are conceived according some relevant historical facts. However, the division in sub-corpora is arbitrary and, from another point of view, could be different. The first sub-corpora include librettos in the period 1600-1690: it spans the Baroque period till the foundation's year of the Accademia dell'Arcadia, which had a strong influence on the libretto's style during the XVIII century (Pavan, 2019, p. 4).

The second sub-corpora span the period 1690-1748. In 1748 Carlo Goldoni started to write librettos in a novel style, also enlarging the lexicon adopted by librettists (Pavan, *ibid.*, p. 5).

The third sub-corpora include librettos from 1748 till 1816. The plots of librettos started to change, including different subjects coming also from the present time. The language of literature in general became closer to modern Italian language. In 1816 the translation of an article by Madame de Staël gave officially the birth to the Romantic age in the Italian peninsula.

The fourth sub-corpora span the years between 1816 and 1850. The language of librettos started to change including more features coming from modern Italian language (Migliorini, 2019, pp. 727-827). In 1850 Richard Wagner finished to write *Lohengrin* and started to travel in Italy to conceive his works. In those years he reformed the drama writing several books about it.

Finally, the last sub-corpora include librettos written from 1850 and beyond. The language of librettists in this period started to use current words coming from the Italian language, while the verses of librettos used less the rhyme schemes, breaking the rules to adapt the text to a more modern music (Fabbri, 1988, pp. 165-170). However, the "golden age" of opera ended in these years, so the number of operas written in this period is smaller.

This historical distribution of texts is just an example on how to proceed with diachronic lexicon's analysis. A different choice of historical periods can generate different results. However, if the sub-corpora take into consideration the changing of language along history, the results of the analysis should be always acceptable.

Results and Discussion

The list of real keywords created with the help of the software CorpStat follows here. It includes the parameters a[b][c] for each sub-corpora. The decimal numbers are truncated for reasons of space, so sometimes the parameter [b] appears to be 0.00, even if in reality it is a smaller number. The keywords validated for only some historical periods are preceded by an asterisk. The rank of the original keywords from AntConc is not changed: the

keywords not validated were simply not printed. Also, some verb forms were avoided in the list. From the original list of 200 keywords the new list includes only 41 validated keywords.

Total amount of words in all sub-corpora: 3209557

Column 1: words

Column 2: 1600-1690: till the Accademia dell'Arcadia foundation --> Total amount of words in the sub-corpus: 804390

Column 3: 1690-1748: till Carlo Goldoni reform --> Total amount of words in the sub-corpus: 627603

Column 4: 1748-1816: till Romantic period --> Total amount of words in the sub-corpus: 757459

Column 5: 1816-1850: till Wagner --> Total amount of words in the sub-corpus: 554298

Column 6: 1850 and beyond --> Total amount of words in the sub-corpus: 465807

1.	core	1046 [0.13][100.0]	790 [0.12][98.07]	714 [0.09][100.0]	558 [0.10][98.36]	272 [0.05][95.34]
2.	pietà	156 [0.01][42.37]	219 [0.03][40.38]	189 [0.02][31.42]	132 [0.02][37.70]	96 [0.02][30.23]
3.	sposa	245 [0.03][72.88]	362 [0.05][86.53]	538 [0.07][88.57]	233 [0.04][80.32]	135 [0.02][79.06]
4.	sposo	164 [0.02][69.49]	412 [0.06][92.30]	454 [0.05][85.71]	194 [0.03][80.32]	53 [0.01][55.81]
5.	*regina	325 [0.04][66.10]	321 [0.05][71.15]	176 [0.02][25.71]	160 [0.02][45.90]	334 [0.07][37.20]
6.	ciel	1088 [0.13][100.0]	353 [0.05][96.15]	536 [0.07][100.0]	720 [0.12][98.36]	359 [0.07][93.02]
7.	figlia	237 [0.02][79.66]	359 [0.05][76.92]	416 [0.05][68.57]	356 [0.06][73.77]	205 [0.04][69.76]
8.	figlio	332 [0.04][74.57]	550 [0.08][61.53]	286 [0.03][52.85]	403 [0.07][63.93]	118 [0.02][58.13]
9.	addio	200 [0.02][62.71]	165 [0.02][76.92]	251 [0.03][81.42]	198 [0.03][88.52]	210 [0.04][95.34]
10.	nume	368 [0.04][89.83]	143 [0.02][75.00]	146 [0.01][57.14]	118 [0.02][65.57]	31 [0.00][30.23]
11.	trono	136 [0.01][57.62]	364 [0.05][75.00]	189 [0.02][42.85]	122 [0.02][50.81]	64 [0.01][39.53]
12.	amore	1407 [0.17][100.0]	767 [0.12][98.07]	776 [0.10][100.0]	399 [0.07][91.80]	397 [0.08][97.67]
13.	gioia	215 [0.02][88.13]	130 [0.02][75.00]	174 [0.02][74.28]	336 [0.06][96.72]	199 [0.04][86.04]
14.	sen	461 [0.05][89.83]	266 [0.04][94.23]	225 [0.02][88.57]	142 [0.02][80.32]	30 [0.00][46.51]
15.	fato	314 [0.03][83.05]	229 [0.03][88.46]	166 [0.02][65.71]	139 [0.02][70.49]	81 [0.01][53.48]
16.	*costanza	108 [0.01][52.54]	283 [0.04][78.84]	266 [0.03][58.57]	55 [0.00][50.8]	23 [0.00][9.30]
17.	amante	626 [0.07][96.61]	485 [0.07][92.30]	315 [0.04][97.14]	124 [0.02][73.77]	67 [0.01][69.76]
18.	cielo	874 [0.10][98.30]	255 [0.04][82.69]	313 [0.04][91.42]	511 [0.09][93.44]	314 [0.06][100.0]
19.	infelice	254 [0.03][89.83]	212 [0.03][84.61]	229 [0.03][81.42]	196 [0.03][83.60]	48 [0.01][60.46]
20.	cara	199 [0.02][84.74]	292 [0.04][100.0]	430 [0.05][100.0]	168 [0.03][77.04]	56 [0.01][62.79]
21.	*reggia	208 [0.02][74.57]	227 [0.03][61.53]	89 [0.01][31.42]	55 [0.00][29.50]	14 [0.00][18.60]
22.	padre	301 [0.03][76.27]	589 [0.09][76.92]	578 [0.07][75.71]	626 [0.11][80.32]	378 [0.08][86.04]
23.	labbro	94 [0.01][50.84]	140 [0.02][65.38]	104 [0.01][65.71]	121 [0.02][83.60]	97 [0.02][83.72]
24.	piè	167 [0.02][38.98]	69 [0.01][34.61]	44 [0.00][27.14]	56 [0.01][31.14]	34 [0.00][27.90]
25.	affetto	221 [0.02][89.83]	220 [0.03][86.53]	215 [0.02][91.42]	143 [0.02][80.32]	68 [0.01][72.09]
26.	empio	263 [0.03][84.74]	231 [0.03][86.53]	136 [0.01][50.00]	101 [0.01][57.37]	36 [0.00][37.20]
27.	guardie	30 [0.00][20.33]	198 [0.03][61.53]	190 [0.02][41.42]	165 [0.02][52.45]	46 [0.00][39.53]
28.	momento	81 [0.01][62.71]	133 [0.02][80.76]	297 [0.03][95.71]	231 [0.04][91.80]	92 [0.01][72.09]
29.	genitor	90 [0.01][42.37]	144 [0.02][59.61]	147 [0.01][55.71]	102 [0.01][63.93]	21 [0.00][23.25]
30.	*beltà	166 [0.02][38.98]	70 [0.01][34.61]	27 [0.00][18.57]	30 [0.00][22.95]	18 [0.00][11.62]
31.	ohimè	188 [0.02][35.59]	14 [0.00][15.38]	59 [0.00][25.71]	27 [0.00][21.31]	19 [0.00][13.95]
32.	innocente	148 [0.01][69.49]	197 [0.03][73.07]	121 [0.01][60.00]	130 [0.02][67.21]	30 [0.00][44.18]

33. caro	271 [0.03][91.52]	367 [0.05][96.15]	520 [0.06][94.28]	176 [0.03][72.13]	57 [0.01][53.48]
34. sol	1001 [0.12][100.0]	575 [0.09][100.0]	535 [0.07][97.14]	438 [0.07][100.0]	289 [0.06][97.67]
35. sorte	455 [0.05][94.91]	444 [0.07][90.38]	391 [0.05][91.42]	200 [0.03][88.52]	102 [0.02][72.09]
36. barbaro	85 [0.01][42.37]	130 [0.02][73.07]	201 [0.02][77.14]	105 [0.01][55.73]	16 [0.00][23.25]
37. istante	35 [0.00][38.98]	53 [0.00][48.07]	162 [0.02][71.42]	188 [0.03][86.88]	94 [0.02][81.39]
38. destin	141 [0.01][76.27]	154 [0.02][78.84]	110 [0.01][57.14]	90 [0.01][72.13]	42 [0.00][44.18]
39. destino	146 [0.01][77.96]	119 [0.01][71.15]	116 [0.01][75.71]	91 [0.01][77.04]	113 [0.02][72.09]
40. signor	504 [0.06][93.22]	462 [0.07][96.15]	878 [0.11][88.57]	368 [0.06][88.52]	232 [0.04][90.69]
41. *bravo	11 [0.00][15.25]	10 [0.00][13.46]	127 [0.01][48.57]	264 [0.04][24.59]	37 [0.00][37.20]

Each word in the list shows a different evolution of use along history. The frequency can change according to the historical period, but this could also depend from the different amount of words in each sub-corpus. For example, the word *sol* shows a decrease of frequency along history, but it must be taken into consideration that also the amount of words in each sub-corpora decrease. It seems more interesting to notice that *sol* has a similar value of the parameter [c] along history. This should mean that the use of *sol* in the librettos did not change substantially in history. The parameter [c] was introduced to avoid that the analysis based on keywords would be criticized on the basis that a word can be a keyword even if it appears only in one text in the corpus. It is known that keywords coming from the Log-Likelihood or Chi-square methods could have this problem (Bestgen, 2017, p.38). However, [c] does not take into consideration the *length* of text in terms of amount of words. If a text in the corpus has a large number of words in comparison with the others, a word which occurs only in that text with a high frequency should be accepted in any case as a keyword. The problem of different length of texts is found also in the libretto's sub-corpora: they have more or less the same number of texts, but the length of texts can be different. However, in the case of librettos, the average length of texts is similar and it is connected with the duration of the operas: with the exception of the Baroque period, in which the duration of the operas could be also several hours, in the other historical periods the duration is quite standardized. To avoid the problem of different texts' length, a simple solution can be at first to check the amount of words of the shortest text in the corpus, using it as a threshold. Then one could break the longer texts, getting shorter texts according to the threshold and later to proceed with corpus' analysis. Something similar is found in the *Brown Corpus*, where each text includes only a maximum of 2000 words (Cresti, Panunzi, 2013, p. 58-59). A fourth parameter could be added to the output of CorpStat, showing the percentage of word's presence according to the minimum length of text. However, this was not the case of librettos, having the texts a similar length in the corpus.

The words preceded by an asterisk do not reach the threshold of [c] = 20 % in all the sub-corpora. Taking into consideration [c], CorpStat shows the evolution of use of the words in the list along history. Among these words the use of *regina*, *costanza*, *reggia*, *beltà* (bellezza), *trono* shows a decrease in history. This should depend mainly from the changing of plots according to historical period. The word *bravo* appears to be used especially in XVIII century, with the same meaning it has today. However, in the XIX century this word could also mean "evil", like it happens in the 1869's libretto *I promessi sposi* by Antonio Ghislanzoni, which derives from the Alessandro Manzoni's work.

Taking into consideration some of the other keywords, the word *core* (cuore) seems to be the most typical in the librettos of all epochs. This word was already important in the Florentine Vernacular literature of XIV century, but it is even more in the librettos' plots. The word *amore* has the highest word frequency in the list and appears almost in all sub-corpora's texts. Nevertheless, its rank is not the highest. The presence of *core*, *amore*, *amante*, *infelice*, *caro* shows the strong connection with the themes of Florentine Vernacular literature. Also, the exclamation *ohimè*, which in Francesco Petrarca's *Canzoniere* appears connected with the themes of *amore cortese*, is found in Florentine Vernacular literature to express "sorrow", like for example in Boccaccio's *Decameron*. Another important theme of the libretto's plots is the mythological one (Bonomi, Buroni, pp. 33-34). In fact, among words, it is possible to find also *fato*, *sorte* and *destino*, which are significant especially in mythological plots: in general,

they express the concept of destiny usually in a tragic situation. *Fato* is more typical in the first historical periods of opera, which were closer to the Greek and Roman myths. *Destino* today replaced *fato* in the spoken language, but this word was popular in all periods. Also, the word *nume* is more typical in Baroque period: this word is usually connected with gods' will and the librettos' mythological plots are more common during the Baroque period. The word *istante* is not common in the major authors of Florentine Vernacular literature. This word is even not mentioned in some editions of *Vocabolario degli Accademici della Crusca*, but it is still possible to find *istantissimo* (VDC, 1686, p. 495). It should be noticed that this word, along history, became more and more used by librettists, especially in the last opera's historical periods. In Italy, at present time, *istante* is currently used also in the spoken language.

Among the most significant words in the list there is also *cielo*, which is used widely in the poetry of all epochs. *Addio* is used much more in the last historical periods: in the Romantic period this word became popular because it was also alluding to the loss of someone or something, being this a recurring theme among romantics.

The word *empio* shows a progressive decrease of use in history and it is used generally with the meaning of "evil" rather than "blasphemous".

Finally, there are words connected to the family status: *genitor*, *padre*, *sposa*, *sposo*. These are typical in the librettos' plots, in which the characters that they represent are often involved.

The few remaining words, like *signor* or *affetto*, seem not to show a particular trend in history. They are also well established in the librettos' plots.

Conclusions

The method proposed in this article allows to observe the occurrence of words along history. In the case of libretto the analysis was performed on a corpus which is rather small (in the history of opera thousands of librettos were written). However, given the fact that opera was a sort of industry, the language and the plots of librettos were quite standardized. Therefore, it is assumed that the results are representative for opera.

The method and the software proposed to analyze lexicon can be extended to a huge variety of linguistic analysis. For example, it is possible to choose shorter historical periods, or to conduct the analysis on many historical periods. In future, it is possible to add more parameters in the output of CorpStat.

Finally, it is also possible to perform the analysis synchronically, getting statistical informations about the lexicon coming from different corpora belonging to the same time period. Further studies will show the many possibilities of this method.

References

- [1] Anthony, L. (2014). AntConc 3.4.3, software. Tokyo: Waseda University.
- [2] Bestgen, Y. (2017). Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora. *Quaderns de Filologia - Estudis lingüistics* 22:33-56. <https://doi.org/10.7203/qf.22.11299>
- [3] Bondi, M., Scott, M., eds. (2010). *Keyness in texts*. Amsterdam: John Benjamins.
- [4] Bonomi, I., Buroni, E. (2010). *Il magnifico parassita. Librettisti, libretti e lingua poetica nella storia dell'opera italiana*. Milano: FrancoAngeli.
- [5] Bonomi, I., Buroni E. (2017). *La lingua dell'opera lirica*. Bologna: il Mulino.
- [6] Brezina, V. (2018). *Statistics in corpus linguistics. A practical guide*. Cambridge: Cambridge University press.
- [7] Cresti, E., Panunzi, A. (2013). *Introduzione ai corpora dell'italiano*. Bologna: Il Mulino.
- [8] Fabbri, P. (1988). *Metro e canto nell'opera italiana*. Torino: EDT.
- [9] Lüdeling, A., Kytö M., eds. (2009). *Corpus linguistics. An International Handbook. Vol. 2*. Berlin: Mouton de Gruyter.
- [10] Kytö, M. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada* v. 11, 2:417-457.
- [11] Migliorini, B. (ed. 2019). *Storia della lingua italiana*. Firenze: Bompiani.
- [12] Partington, A. (2011). Corpus Linguistics: what it is and what it can do. *Cultus – the Journal of Intercultural Mediation and Communication* 4:35-58.

- [13] Pavan, L. (2019). Some Language Features in Italian Opera Librettos of XVII-XVIII Centuries. *International Journal of Literature and Arts*, Vol. 7, 6:172-178, doi: 10.11648/j.ijla.20190706.17
- [14] Staffieri, G. (2014). *L'opera italiana. Dalle origini alle riforme del secolo dei Lumi (1590-1790)*. Roma: Carocci.
- [15] VDC – *Vocabolario degli Accademici della Crusca*. Venezia: Per Combi, e La Noù, 1686.