Textometric methods and the TXM platform for corpus analysis and visual presentation

UDC 811.163.41'322
DOI 10.18485/infotheca.2019.19.1.2

Jelena Jaćimović
jelena.jacimovic@stomf.bg.ac.rs

University of Belgrade
School of Dental Medicine
Belgrade, Serbia

ABSTRACT: Textometric approach has long been applied as a useful method for corpus analysis in various fields of humanities and social sciences. Textometry allows the non-linear quantitative and qualitative study of digital corpora, combining lexicometric and statistical research with developed corpus technologies. In this paper, the current version of the srpELTeC corpus was analyzed within the TXM program environment to illustrate the possibilities of the textometric approach and visual presentation of the obtained results.
KEYWORDS: textometry, digital corpora, Serbian language, TXM, srpELTeC.

PAPER SUBMITTED: 29 June 2019
PAPER ACCEPTED: 6 September 2019

1 Introduction

Digital corpora serve as valuable and practical sources of empirical data necessary for linguistics and other humanities and social sciences research. With the development of the digital era and the achievements of language technology use, the traditional way of accessing the text changes, opening new possibilities for analyzing large amounts of textual data using statistical methods. Unlike conventional linear reading (close reading), distant reading enables an understanding of literature by collecting and analyzing large amounts of data. Textometry stands out as one of the disciplines which allows a different way of "reading" the texts. Textometry enables the non-linear quantitative and qualitative study of digital corpora, combining lexicometric and statistical research with developed corpus technologies.

1.1 Textometry

The beginnings of textometric research relate to France and the work of Pierre Giraud (1954; 1959) and Charles Muller (1973), who dealt with problems and methods of linguistic statistics. The methods developed by Jean-Paul Banzécri and his colleagues and students, also applied to linguistic data (Benzécri, 1973), are adopted and implemented in textometry. The methodological basis of textometry could be also found in the works of Ludovic Lebart and André Salem (Lebart and Salem, 1988, 1994). In addition to the methods adopted, textometry has developed new statistical models to discover important features of textual data, such as contextual „attractiveness" of the words, the linearity and internal text structure, intertextual contrasts or the indicators of lexical evolution (characteristic period of a word usage, detection of the significant usage disruptions). The textometry analysis results give a synthetic, selective, and suggestive overview of the re-organized text, seen now through the hierarchical lists, visual maps, re-grouping, and text enhancements. The new way of accessing versatile, controlled textual data and a new way of „reading" the text based on the data that were previously not available, highlights the heuristic power of statistical tools in the analysis of literary texts.

A textometric approach implies that each text has its internal structure, which would be difficult to analyze manually. Facilitated by computer tools and created hypertext links, textometry based on numerical indicators simultaneously provides a general synthetic view of the text, as well as the possibility of local insight into the context. This text „closeness" during analysis, as well as a balanced approach to both the general and the local in the text, opens a range of hermeneutical questions and reveals a linguistic reality that is a highly significant representing valuable observational field. However, it should be kept in mind that textometry is a process that provides results and the identification of patterns and trends that would otherwise remain hidden due to large amounts of data, but that the interpretation of the obtained results and its validity depend on the experts and the system used for textometric analysis.

Beside textometry, other disciplines also use quantitative approaches for the analysis of textual data. Information retrieval (IR) deals with finding robust methods for managing large quantities of texts too. Nevertheless, IR focuses on finding and linking certain documents and information discovered in them. In contrast to the field of IR, textometry applies to a closed, stable corpus of texts. Another area that can compare with textometry in a cer-

tain sense is Latent semantic analysis (LSA). LSA also uses mathematical methods for the text analysis, but to investigate some out of the text fields, such as language and other cognitive functions. Quantitative methods that are known and applied in textometry are also used, for example, in the field of Text mining (TM). Unlike TM's basic idea to find and extract remarkably valuable information, the textometric approach focuses on the text itself and the discovery of linguistic trends and rules, their text realizations and changes. Furthermore, the subjects of textometric analysis are well-known corpora, documents, and objects. Natural language processing (NLP) also uses statistics for the system building and recognizing the linguistic units of the text. Although the objectives of these two areas differ, arguably they complement one another. For example, corpora can be used in the NLP to set up a system designed to identify specific entities or to build recognition rules (such as term extraction or morphological and syntax tagging), which can be of great importance in the later textometric analysis' process. On the other hand, the textometric approach to the corpus research provides the possibility of recognizing some entities or the text specificities, as well as text tagging, useful for the NLP tools development.

1.2 The TXM programming environment

The quantitative approach to researching textual corpus elements is not new, but it is significantly simplified and improved using existing tools. The programs designed to analyze large corpora do not lead to the discovery of new language information but offer a novel perspective on the perception of the already known (Hunston, 2002). When analyzing large corpora, it is essential to allow users to run multiple different queries and navigate through text in an easy-to-use environment. Several software solutions, which are free to use and have developed graphical user interface (Pincemin, 2018), allow text data analysis using the textometry primary functions (such as the concordance view, the *specificity score*, or correspondence factor analysis discussed later).

Based on textometry, a methodology that allows quantitative and qualitative analysis of textual corpora, TXM (Heiden et al., 2010; Heiden, 2010) is an open-source program,[1] widely used in research in various fields of social sciences (history, literature, geography, linguistics, sociology, political sci-

[1] Desktop installations for Windows, Linux, and Mac OS X are available, as well as a web platform.

ences). The graphical user environment of the TXM uses the CQP[2] (Corpus Query Processor) browser and the R[3] statistical package. The TXM allows the study of abundance of any language materials, a large number of input text formats, the use of various tools for processing natural language inputs (such as automated lemmatizers). This program enables the construction of a sub-corpora or parts based on metadata (date, author, genre, etc.) or corpus structural units (like text, chapter, paragraph), querying (using the CQP browser), and more complex query results processing based on quantitative methods, as well as the export of results in a tabular or graphical form.

Digital corpora possible to analyze within this environment are written texts, transcripts (synchronized with original audio or video), and parallel corpora. The TXM allows the import of texts encoded under certain conventions, such as texts in the recommended UTF-8 (TXT format) or XML[4] (eXtensible Markup Language) documents encoded following TEI[5] (Text Encoding Initiative) instructions (XML or XML-TEI format). Hence, it is possible to select, within the TXM, a representation level of corpus texts that is more or less rich and therefore more or less demanding for preparation. The plain text representation offers some essential analysis options. On the other hand, due to the ampler text representation and its internal structure, XML-TEI texts can be explored in far more detail (Lavrentiev et al., 2013).

During the import, TXM generates a customized version of the TEI data model - an XML-TXM format based on source texts and formats, used as a base model for all analyzes. This corpus representation implies the existence of specific corpus units, such as textual, structural, and lexical units. Textual units are **texts** that the corpus comprises (such as books, articles, interviews) and they can have their attributes or metadata (like author, title, date, genre). Then, each of the texts can have a specific structure and several internal **structural units** (for instance, chapters, passages, administrative speech) that may have particular properties or attributes (such as an address, number). Lastly, **lexical units** are defined, because each text is composed of a series of words that can have specific properties, such as graphic form, lemma, or grammatical category. Metadata, like all XML text elements, is considered within the TXM as a structural unit. Thus, the pos-

[2] CQP (on-line)
[3] R (on-line)
[4] XML (on-line)
[5] TEI (on-line)

sibilities of text analysis depend on its representation. Existing text/corpus annotations, along with structural and lexical units, are used to create a sub-corpus or corpus parts to compare them and search. Therefore, from the corpus research standpoint, it is necessary to define in more detail the units used for the analysis. The TXM builds an HTML format for each corpus textual unit, providing the possibility of returning to the text at any analysis phase.

Although TXM text import modules provide automatic morphological and syntactic tagging, as well as lemmatization of texts using the TreeTagger (Schmid, 1994), it is possible to preprocess corpus data beyond the platform using other NLP tools. There is no standard representation of the results of these tools, and only the standards used in practice apply. The results of NLP tools TXM recognizes as annotations added to the XML-TEI text representation.

In addition to built XML-TXM format, the CWB[6] (Corpus Workbench) format is also generated, applied to search the corpora using queries expressed by CQP syntax. A description of the CWB environment and regular expressions, used by Corpus Query Language (CQL), can be found in (Utvić, 2014; Evert and Hardie, 2011).

The TXM environment primarily enables qualitative text analysis through generated frequency lists, concordances, or the HTML text edition. Any combination of the properties of defined text units can be used to query and display the contexts in which those units appear. On the other hand, statistical models, counting the properties of lexical units, permit quantitative analysis, i.e. analysis of their corpora distribution (factor analysis, cluster analysis), their remarkably high or low representation in certain parts (specificity analysis), or the analysis of the lexical attraction between words (co-occurrence analysis). Each result of the analysis can be exported for further examination and editing, using another tool, in tabular or graphical form.

This paper aims to present the current version of the srpELTeC corpus[7] and to illustrate the possibilities of the textometric approach and the application of the TXM tools for analysis and visual presentation of the results. The conducted analysis of the Serbian novel corpus from the late 19th and early 20th centuries should highlight the potential of textometry and bring

[6] CWB (on-line)
[7] srpELTeC (on-line)

it closer to researchers from the various scientific fields involved in corpus analysis.

2 The methodology of the srpELTeC corpus textometric analysis

2.1 The srpELTeC corpus

The corpus used for this paper is called the srpELTeC corpus. The main motive for the creation of this corpus is its inclusion in a multilingual European Literary Text Collection, which should contain 100 novels for each of the languages included in the COST Action *Distant Reading for European Literary History*, published between 1850-1920 that have expired copyright.[8]

The Serbian corpus, unlike many other European languages involved in this action, is produced from the very beginning. Most Serbian novels from this period were not digitized or properly digitized, especially since the first editions of many novels were hard to obtain. Serbian literature, and especially the Serbian novel, can by no means be compared in scope with the literary „production" of the major European languages, such as French, English, or German. Therefore, the novel selection and finding printed copies were extremely demanding. Transformation into a machine-readable form involved scanning and optical character recognition (OCR). OCR errors were automatically corrected using a specialized tool based on the Serbian morphological dictionary (Krstev, 2008). A large number of volunteers were engaged in manual correction of the remaining errors and structural annotation markup.[9] At this stage, following the requirements of the COST Action, the metadata was also prepared to be used for later corpus analysis.

It is well-known that corpus size, its representativeness, and balance should be taken into account when designing a corpus (O'Keeffe and McCarthy, 2010). The preparation of several novels published in the Serbian

[8] The ELTeC collection should contain the first editions of literary texts (novels) from a distinct period and written in several languages. To be included in one of the ELTeC sub-collections, the text must have been first published as a book (minimum length of 10.000 words) in a European country between 1850 and 1920. Other novel selection criteria primarily concern the author's gender and the canon. Each ELTeC sub-collection should contain at least 10% to 50% of texts written by female authors, as well as at least 30% of both prestigious (highly canonized, reprinted more than once) and unknown (not or once reprinted) novels.

[9] Manual correction (on-line)

language in the period 1850-1920 is in progress, and for the time being, 21 works have been included in the ELTeC corpus, digitized until the writing of this paper (Table 1). The reason for the selection of these works, therefore, is neither an aesthetic nor a thematic nature. Since the SerbianELTeC corpus currently does not include all the novels published in a given period, it cannot be said to be representative or balanced. Besides, for example, most of the works included were written by male authors. However, this paper aims to demonstrate the implementation of textometric analysis using the TXM tool for which the created Serbian literature corpus from the late 19th and early 20th centuries can be a good source. Furthermore, this specialized corpus contains a collection of texts of exceptional significance that are not an example of a modern language and in which, besides well-known authors and their works, there are those whose work brings the beginnings of a modern novel structure or those about which is written insufficiently in the history of Serbian literature. For example, the corpus includes novels of Milutin Uskoković, whom critics and historians of literature considered as the originator of Belgrade's, urban style, but also a novel by little-known author Dragomir Šišković. Moreover, the first Serbian science-fiction novel *Jedna ugašena zvezda* by Lazar Komarčić is part of the srpELTeC corpus too, as well as the novel *Babadevojka* by Draga Gavrilović, the first female author who wrote the novel in the Serbian patriarchal society of the time. The current version of the srpELTeC corpus is available in the ELTeC collection.[10]

The texts of the srpELTeC corpus are encoded in XML format, very useful for later analysis, following the TEI guidelines. The header of the TEI document contains bibliographic information about the electronic and original version of the novel, as well as information about the persons responsible for the particular phases of creating and updating the electronic version. Documents are structurally annotated, containing information on the logical structure of the text. In addition to the recommended TEI elements and attributes for structural text annotation (like header, text, body, unit of text - chapter, title and subheadings, paragraph, quotation, words or sentences written in the language other than the language of the text), metadata in the form of a CSV document also includes supplementary information about the gender of the author and the publication type (like novel, story or short prose).

[10] ELTeC collection (on-line)

Author	Publication	Year	Length (w)
Gavrilović, Draga	Babadevojka	1887	23.858
Gavrilović, Andra	Prve žrtve	1893	44.929
Kostić, Tadija	Gospoda seljaci	1896	39.349
Mijatović, Čedomilj	Rajko od Rasine	1892	50.305
	Ikonija, vezirova majka	1891	28.332
Milićević, Milan	Deset para	1881	12.365
	Jurumusa i Fatima	1879	21.947
Stanković, Borisav	Uvela ruža	1899	12.748
	Pokojnikova žena	1902	12.701
Šišković, Dragomir	Jedan od mnogih - roman iz prestoničkog života	1920	21.676
Uskoković, Milutin	Potrošene reči	1911	14.580
	Došljaci	1910	97.467
	Čedomir Ilić	1914	65.073
Dimitrijević, Jelena	Nove	1912	116.782
Ilić, Dragutin	Hadži Đera	1904	65.554
Janković, Milica	Kaluđer iz Rusije*	1919	8.279
Komarčić, Lazar	Dragocena ogrlica	1880	65.160
	Jedna ugašena zvezda	1902	58.334
	Prosioci	1905	28.327
Nušić, Branislav	Opštinsko dete	1902	77.994
Sekulić, Isidora	Đakon Bogorodičine crkve	1919	62.414

* This publication will not be included in the srpELTeC corpus because of its length

Table 1. Literary works included in the srpELTeC corpus used for textometric analysis

The TXM environment recognizes defined metadata as a new structural element `text`, represented by the following attributes: `author`, `title`, `date`, `gender` and `type`. Thus, the distribution of the prepared texts was determined by the author's name, the title, year of publication, the author's gender, and the publication type. Metadata was used in the TXM environment to split the corpus into parts, create a sub-corpora, and for text search.

For the analysis of the srpELTeC corpus, a collection of texts in XML format was imported into the TXM environment using the XML-TEI Zero + CSV import module. Tagging of the srpELTeC corpus texts was done using the TreeTagger and a linguistic model developed for the Serbian language (Utvić, 2011). Automatic segmentation, tokenization, lemmatization, and Part of Speech (PoS) tagging were performed during the corpus import. Within the TXM environment the results of the tagger are treated as lexical units, namely: `n` (numerical position of the word form in the corpus), `srlemma` (the lemma associated with the token by automatic annotation using the TreeTagger program), `srpos` (PoS associated with the token by automatic annotation using the TreeTagger program) and `word` (concrete token realization in the text) (Example 1).

Example 1. *... из нашега* ***друштвеног*** *живота ...* 'from our **social** life' (part of the text from Lazar Komarčić's novel *Jedna ugašena zvezda*)
`n`: 52.726
`srlemma`: друштвен
`srpos`: A
`word`: друштвеног

The import module generates an XML-TXM format based on the XML-TEI text representation, used as the basic model to represent corpus annotations in the TXM. In addition to building the XML-TXM format, the conversion and generation of the CWB format, used for corpus search applying CQP queries, were also performed. The corpus analysis methodology that the TXM environment enables will be described in the next section.

2.2 Textometric methods in the TXM environment

The frequency is the essential parameter within a corpus, indicating how many times a lexical unit appears in a particular corpus context (Dobrić, 2009). Frequency data serves as the basis for conducting various statistical analyzes, providing an empirical foundation for deriving theories about a language phenomenon.

The primary corpus method is the production of frequency lists. Comparison of absolute frequencies of lexical units (exact number of occurrences in the corpus) can be useful and gives an initial impression of the contrast that exists among the corpus parts, but for the comparison of frequencies in parts of different sizes, it is necessary to normalize, or express the frequencies by a common factor – relative frequency. It would be expected that the relative frequency is calculated as the ratio of the absolute frequency of

the lexical unit and the total number of units in the part of the corpus. The calculated mean value is a mathematical expectation for the normal (Gaussian) distribution of probability. However, the appearance of lexical units in some parts of the corpus is not necessarily consistent with the normal distribution. Pierre Lafon (Lafon, 1984) noticed that the probability of occurrence of lexical units is consistent with the hypergeometric (negative binomial) distribution. The probability that lexical unit A, which is part of corpus vocabulary V, will occur f times in corpus part p of length t, taking into account the total number of occurrences of this unit F in the whole corpus of length T, proposed in (Lafon, 1980), is calculated by the formula:

$$Prob_{specif}(card\{A \in V | A \in p\} = f) = \frac{C_F^f \times C_{T-F}^{t-f}}{C_T^t}, \text{ where}$$
$$C_n^k = \frac{n!}{k!(n-k)!}$$
$$n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$$

The calculation of the *specificity score* based on the hypergeometric distribution in the TXM environment shows the probability of a lexical unit occurring in a particular part of the corpus. The TXM also provides a graphical representation of the specificity distribution of the selected units. *Specificity score* values higher (positive) or lower (negative) than expected express a more or less represented lexical unit. Thus, it is possible to identify significantly common (positive keywords) or significantly rare (negative keywords) occurrences of a lexical unit in parts compared to the whole corpus, which is a useful starting point for making assumptions about text keywords, domain, or authorship.

Besides the specific frequencies, another standard textometry method is a correspondence factor analysis (Benzécri, 1973). The principles of correspondence factor analysis, developed within the French school „Analyse des Données", have been used to analyze the corpus, but also many other data types (Beaudouin, 2016). This statistical technique enables the display and review of the data set, that is, the interdependencies that exist between corpus and lexical units, in a two-dimensional graphical form.

The initial idea was to discover the patterns of interrelations between two sets of elements recorded in a tabular form. For the sample of textual corpora, if the table columns and raws contain texts and words respectively, at the intersection of columns and rows there are indicators of the presence or word frequency in the text (like word frequency, the specific word frequency). The information contained in the matrices can be synthesized using the

data analysis algorithms. Factor analysis aims to re-organize the matrices containing the maximum amount of information. In other words, the basic idea of conducting correspondence factor analysis is to simplify a complex data set (data cloud) and to find ways to present as much information as possible in a smaller space. Firstly, the gravity center and dispersion of the cloud is calculated. The factor planes and the main axes of the dispersion are constructed in the following step. The points are projected on these planes, and their coordinates on these axes are factors. The plane defined by the first two axes can produce the best cloud projection, which minimizes the loss of information. The main goal is to visualize the distance between the attributes, that is, from the random distribution.

Correspondence factor analysis is often combined with cluster analysis, i.e. hierarchical classification based on the coordinates of the factor axes elements. This classification method serves to identify homogeneous subgroups of texts and words. Cluster analysis, applied along with factor analysis, enables a better understanding of the data and simplifies its interpretation.

3 The results of the textometric analysis

3.1 Corpus general information and frequency

The observed srpELTeC corpus includes texts containing a total of 935.902 words. Regarding the representation of the texts of a particular author, the most extensive parts of the corpus are the texts of Milutin Uskoković, Lazar Komarčić, and Jelena Dimitrijević, while the part containing Milica Janković's novel includes only 8.279 words (Figure 1). The results show that in this corpus, which has 78.542 tokens, 32.604 lemmas were used. The most frequent 20 tokens, carrying little semantic information, account for almost 30% of the total number of concrete realizations in the corpus, while 42.004 tokens appear only once in the corpus (hapax), making slightly more than 50% of the total number of tokens. Table 2 shows the different words of the srpELTeC corpus (`word`), their exact number of occurrences in the corpus (absolute frequency F) and rank in the frequency list sorted by descending frequency (rank). The registered most common words, as well as in the case of the Corpus of Contemporary Serbian Language (SrpKor) (Utvić, 2014), are functional words from closed classes of words such as prepositions, conjunctions, auxiliary verbs, or pronouns.

A frequency list of specific word types is generated based on PoS tags. Table 3 shows the values of the `srpos` attributes, their absolute frequency

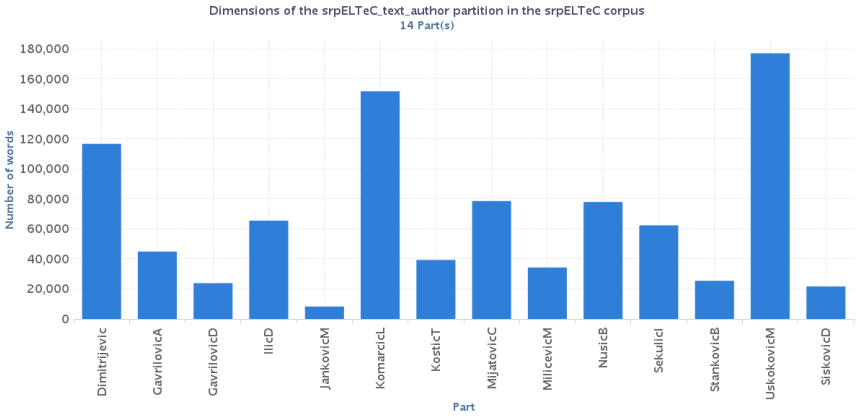

Figure 1. Dimensions of the srpELTeC corpus parts created based on authorship

.

(F) and rank in the frequency list sorted by descending frequency (rank). Aside from nouns and verbs that dominate the texts of the srpELTeC corpus, pronouns also emerge as a morphological category, whose complexity of use and expressive possibilities would be interesting to explore with the TXM tool. Considering the fact that the use of the personal pronoun is arbitrary in the Serbian language because the given verb form also indicates the person, which characterizes the neutral expression style, the high frequency of the pronouns is inherent in stylistically specific literary texts, such as the texts of the srpELTeC corpus (Katnić-Bakaršić, 1999).

Based on frequency data, the TXM allows visual representation of the frequency of particular linguistic phenomena throughout the corpus or parts thereof created based on the existing structural units. Hence, it is easy to notice parts of the corpus and frequency of particular words or expressions. An illustrative representation of the frequency and position of using the lemmas *васељена* 'universe', *црква* 'church', *љубав* 'love' and *девојка* 'girl' in the texts of the entire srpELTeC corpus is given in Figure 2. For example, the lemma *љубав* 'love' is most often mentioned in the novels *Babadevojka*, *Došljaci* and *Đakon Bogorodičine crkve*, in which love is one of the dominant motifs, while the significant use of the lemma *васељена* 'universe' is observed exclusively in the first Serbian science-fiction novel *Jedna ugašena zvezda* of Lazar Komarčić. The lemma *црква* 'church' is mostly used by Nušić at

srpELTeC			SrpKor		
rank	word	F	rank	word	F
1	и	28.545	1	и	4.330.865
2	је	25.422	2	је	4.103.542
3	се	21.128	3	у	3.513.009
4	да	18.932	4	да	3.261.285
5	у	14.932	5	се	2.107.336
6	на	9.233	6	на	1.751.270
7	не	6.721	7	за	1.381.402
8	а	5.642	8	су	1.258.361
9	од	5.234	9	од	919.922
10	што	5.011	10	са	779.469
11	су	4.935	11	а	740.476
12	као	4.857	12	који	650.144
13	за	4.460	13	не	612.218
14	па	4.253	14	о	517.105
15	то	4.132	15	ће	505.643

Table 2. The first 15 rows of the srpELTeC and SrpKor corpora frequency lists

the beginning of his work *Opštinsko dete*, unlike Isidora Sekulić's Đakon Bogorodičine crkve, where this lemma is evenly mentioned throughout the entire novel.

3.2 Specificity score

The frequency of nouns, adjectives, verbs, and adverbs in the Serbian ELTeC corpus is shown in Table 3. Their frequency in the texts of a particular author (f), along with the *specificity score* (S) of the given word type for the entire corpus, is shown in Table 4. The frequency distribution of these types of words in the srpELTeC corpus, partitioned based on authorship, is illustrated in Figure 3. The results reveal that the adjectives are extremely specific for the texts of Lazar Komarčić (S_A=205, 6), whereas Tadija Kostić and Andra Gavrilović use nouns much more often than other authors whose works are included in this corpus (S_N=162, 7 and S_N=83, 6, respectively). On the other hand, the verbs are far less used by Lazar Komarčić compared to their degree of use throughout the corpus, which is presented by the high negative value of the *specificity score* (S_V=−104, 4). The particularly

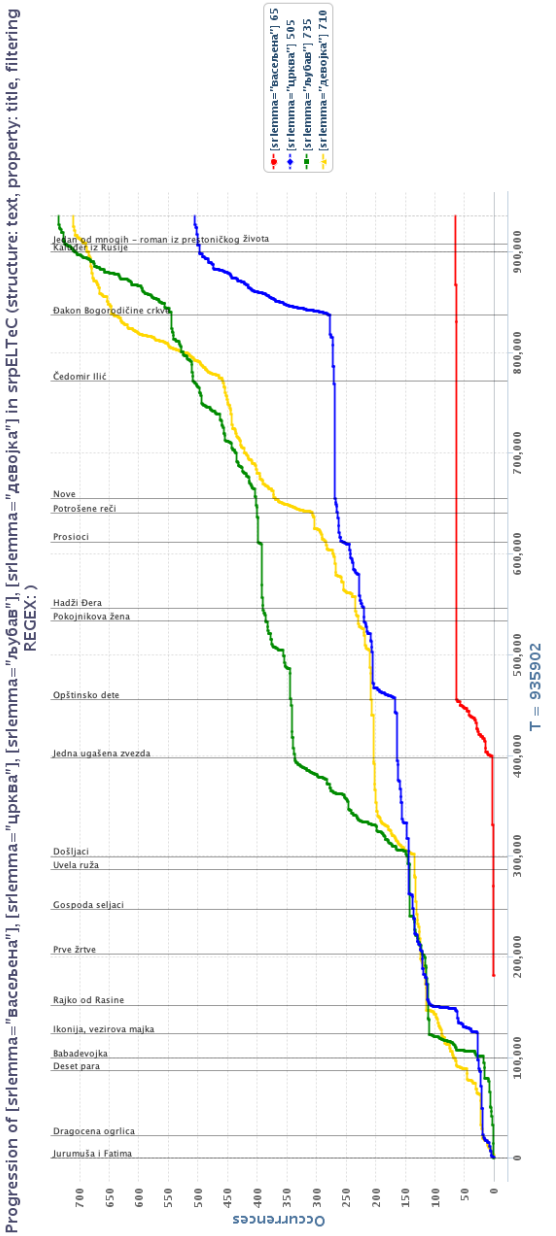

Figure 2. Graphic representation of the specific lemmas' use in the srpELTeC corpus texts

rank	srpos	F
1	N (noun)	192.865
2	V (verb)	173.994
3	PUNCT (punctuation)	103.824
4	PRO (pronoun)	97.239
5	CONJ (conjunction)	90.248
6	A (adjective)	64.242
7	PREP (preposition)	62.584
8	ADV (adverb)	50.628
9	SENT (sentence end marker)	49.184
10	PAR (particle)	36.307
11	NUM (number)	9.208
12	UNDEF (undefined)	2.272
13	? (non-Serbian words or suffixes in compounds)	2.033
14	INT (interjection)	680
15	ABB (abbreviation)	527
16	RN (Roman numeral)	46
17	PREF (prefix)	21

Table 3. Frequency list of position attribute `srpos` possible values in the srpELTeC corpus

low frequency of adjectives is observed in the novels of Branislav Nušić and Dragutin Ilić (S_A=−120,5 and S_A=−82,4), while the nouns, given their degree of use in other parts of the corpus, are far less represented in the stories of Borisav Stanković (S_N=−108).

3.3 Correspondence factor analysis and cluster analysis

In order to simplify the presentation and provide better visibility of the obtained correspondence factor analysis results, the srpELTeC corpus is divided into four parts only (which is the minimal number of corpus parts over which a correspondence factor analysis can be carried out), based on the gender of the author and the century in which his work was published. Therefore, the following corpus parts are marked: *f19* – works of the female authors, published in the 19$^{\text{th}}$ century; *f20* – works by the female authors, published in the 20$^{\text{th}}$ century; *m19* – works of the male authors, published in the 19$^{\text{th}}$ century; and *m20* – works of the male authors, published in the

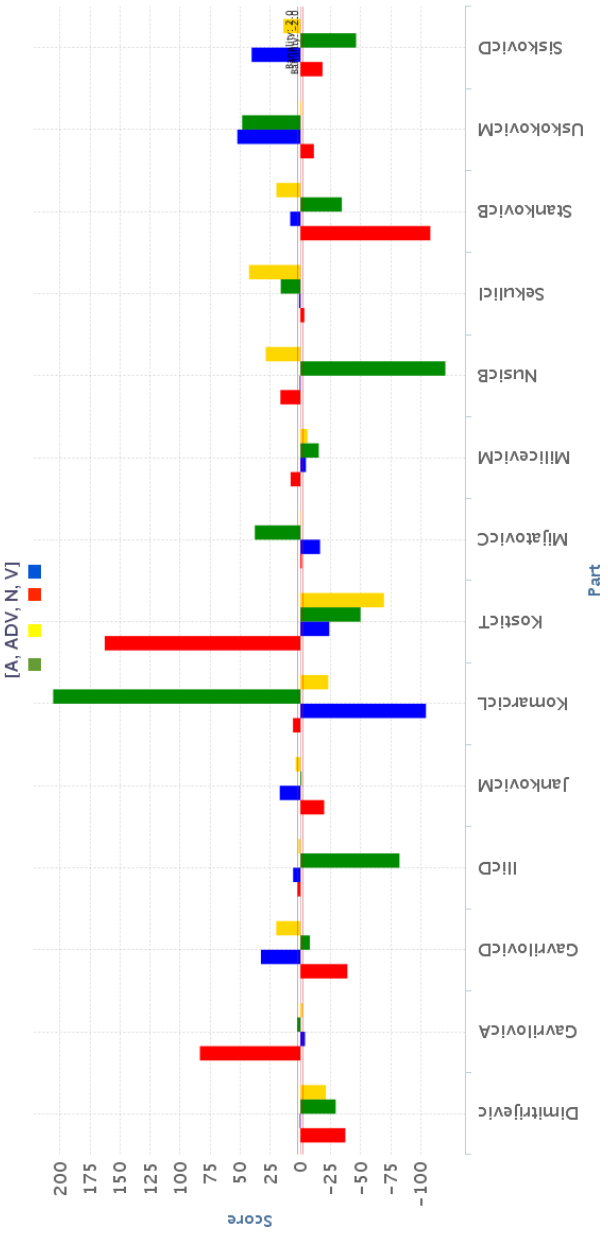

Figure 3. The specificity of nouns (N), verbs (V), adjectives (A) and adverbs (ADV) use in the srpELTeC corpus by authors

Author	f_N	S_N	f_V	S_V	f_A	S_A	f_{ADV}	S_{ADV}
Uskokovic M	35331	-11.2	35446	52.5	13502	48.4	9577	-0.8
Komarcic L	31869	6.2	25443	-104.4	13193	205.6	7481	-23.1
Dimitrijevic J	22330	-37.4	21938	0.5	7066	-29.3	5687	-21.2
Nusic B	16926	16.6	14675	0.6	3812	-120.5	4952	28.9
Mijatovic C	15985	-1.2	13862	-16.3	6259	37.9	4276	-0.4
Ilic D	13728	2.4	12740	6.1	3319	-82.4	3684	1.6
Sekulic I	12497	-3.3	11826	1.1	4772	16.4	4183	42.7
Gavrilovic A	10879	83.6	8125	-3.8	3215	2.7	2356	-1.7
Kostic T	10276	162.7	6606	-24.0	1983	-50.0	1411	-69.5
Milicevic M	7464	8.1	6140	-4.6	1981	-15.3	1680	-5.9
Gavrilovic D	4105	-39.0	5197	32.8	1417	-7.9	1635	20.1
Stankovic B	3867	-108.0	5126	8.5	1268	-34.4	1731	20.0
Siskovic D	3937	-18.4	4838	40.6	981	-46.3	1445	14.1
Jankovic M	1371	-19.8	1860	17.2	539	-0.9	530	4.0

Table 4. The frequency of nouns (N), verbs (V), adjectives (A) and adverbs (ADV) by authors and their *specificity score* for the whole corpus

20$^{\text{th}}$ century. The size of the corpus texts, depending on the gender of the author and the century in which the work was published, is shown in Figure 4, where it can be seen that the part containing the works of the 19$^{\text{th}}$-century female authors is significantly lower than the other parts.

Data on the frequency of nouns, adjectives, verbs, and adverbs in the srpELTeC corpus, divided into parts based on the gender of the author and the century in which the work was published, are shown in Table 5. For each word type, the total number of occurrences in the whole corpus (F), the total number of occurrences in the texts of a particular part (f) and the *specificity score* (S) of the given word type with respect to the entire corpus, are given.

The specific use of word types characteristic for the male or female authors and period (19$^{\text{th}}$ or 20$^{\text{th}}$ century) is also presented in Figure 5. In works published by the 19$^{\text{th}}$-century male authors (part *m19*) nouns are far more prominent than in other parts (S_{m19}=129, 4128), while the use of the verbs is more specific for the part *f19* (S_{f19}=32, 8464), consisting of Draga Gavrilović's novel published in 1887. The specificity of the verbs and adverbs usage is statistically significantly negative in works published by male

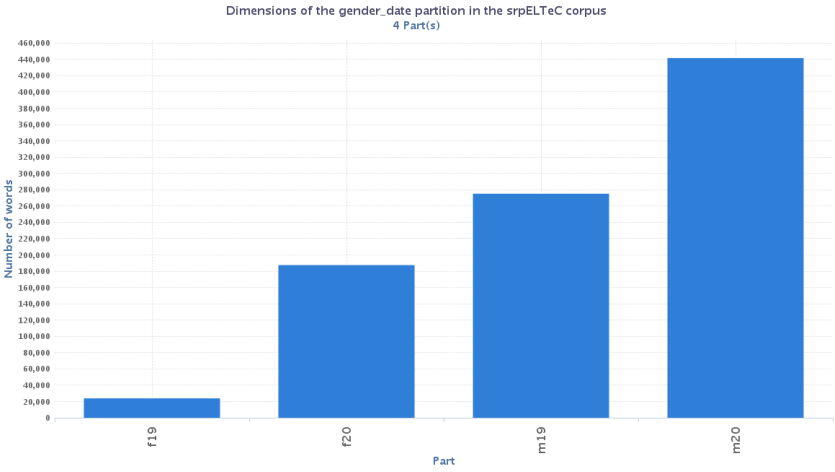

Figure 4. Dimensions of the srpELTeC corpus partitioned based on author's gender and the century in which the work was published

authors in the 19$^{\text{th}}$ century (S_{m19}=−48, 6432 for verbs, S_{m19}=−53, 5126 for adverbs).

Based on the frequency of word types, and according to the χ^2 distribution, correspondence factor analysis was carried out and presented in a two-dimensional graphical form (Figure 6). The obtained factorial map shows that verbs and adverbs are more commonly used in the parts *f19* and *m20*, so they are positioned on the same side of the horizontal axis (the *specificity score* has a positive value). On the opposite side is the part *m19*, which has a markedly negative *specificity score* of the use of verbs and adverbs, but also the part *f20*, whose *specificity score* of the use of verbs and adverbs is positive, but indicates a much lower use of the verbs and adverbs in this part in relation to parts *f19* and *m20*. For this reason, the parts *m19* and *f20*, characterized by a smaller representation of verbs and adverbs, are in opposite quadrants of the vertical axis. Looking at the vertical axis, we can see that there is a part *m19* on one side, in which an extremely high *specificity score* for the nouns is recorded, while the parts in which the nouns are far less represented are located on the opposite side.

The visual representation of the results of the correspondence factor analysis conducted over the corpus divided into author-based parts, and in terms

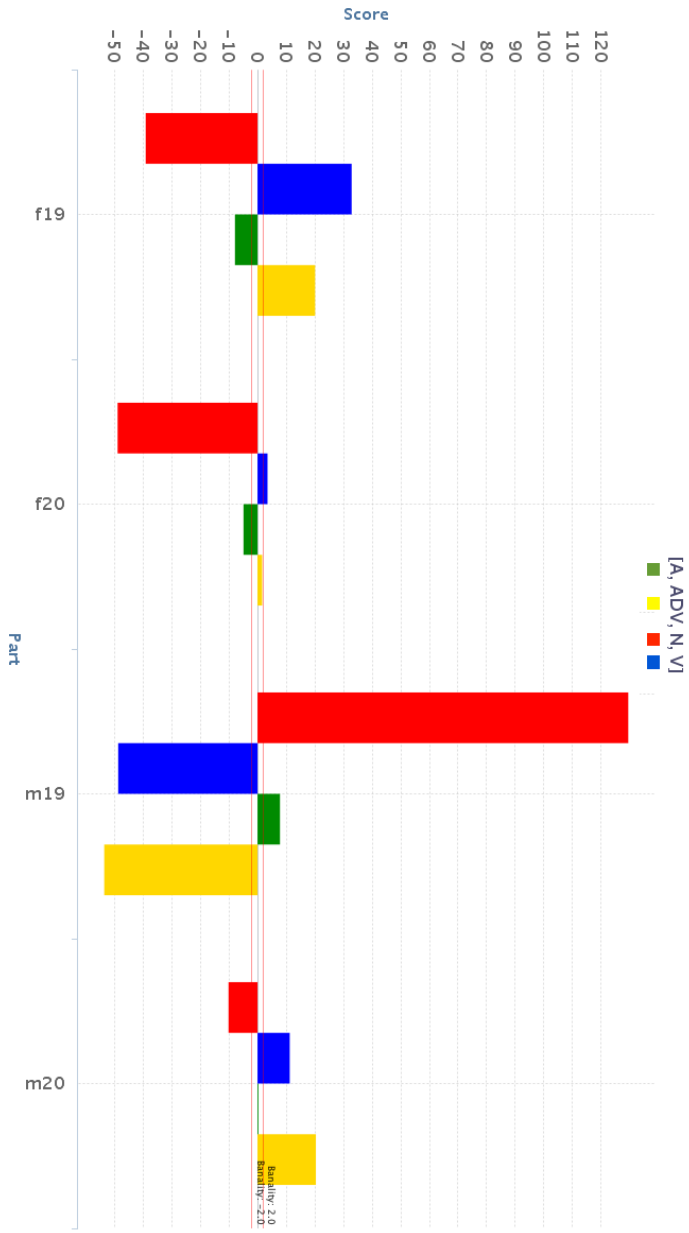

Figure 5. The specificity of certain word types in the srpELTeC corpus by author's gender (m – male or f – female) and the period (19th or 20th century)

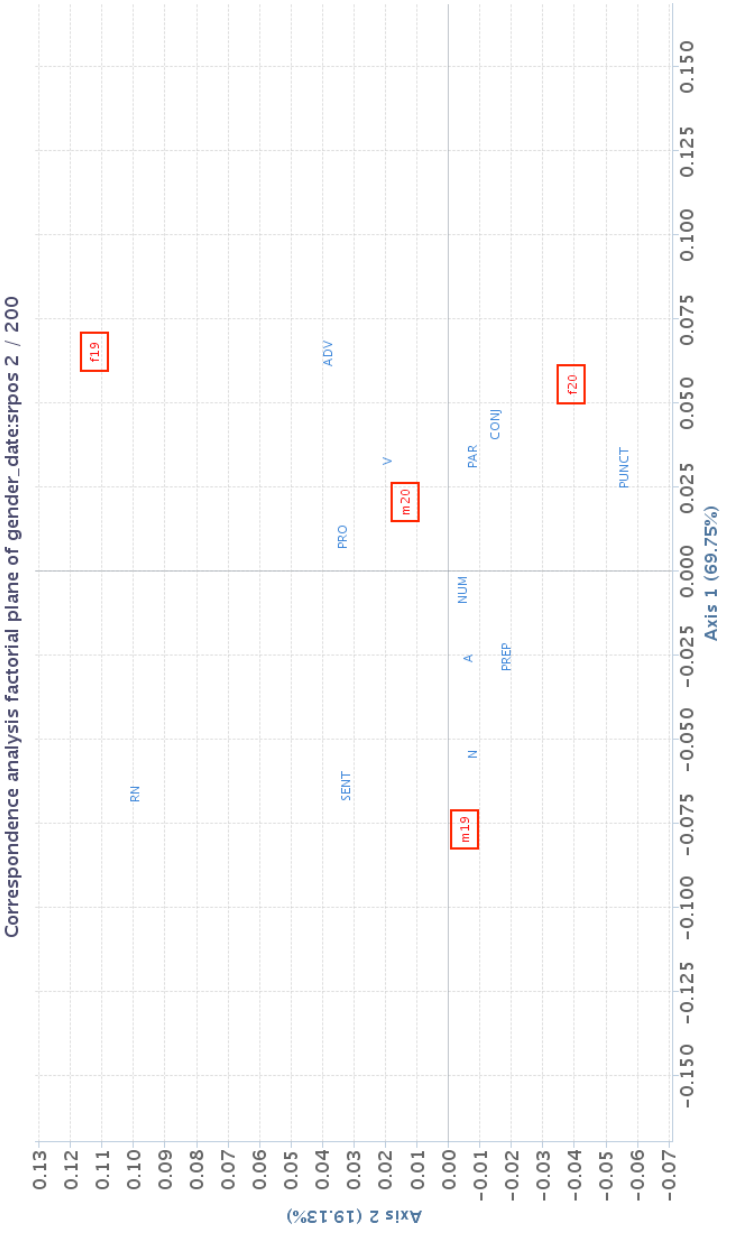

Figure 6. The result of the correspondence factor analysis applied over the corpus partitioned based on author's gender and the publication date

Unit	F	f_{f19}	S_{f19}	f_{f20}	S_{f20}	f_{m19}	S_{m19}	f_{m20}	S_{m20}
N	190565	4105	-39.0398	36198	-48.8455	60820	129.4128	89442	-10.1284
V	173822	5197	32.8464	35624	3.4805	49007	-48.6432	83994	11.2368
A	63307	1417	-7.8845	12377	-4.9007	19383	7.8305	30130	0.3083
ADV	50628	1635	20.0939	10400	1.6149	13477	-53.5126	25116	20.3649

Table 5. The frequency and the *specificity score* of nouns (N), verbs (V), adjectives (A) and adverbs (ADV) by parts created based on author's gender and the period

of the specificity of the used lemmas, is far more complex (Figure 7). The analysis conducted in this way enables the identification and study of laws and trends, otherwise not easily noticeable due to a large amount of diverse data. For example, Milutin Uskoković and Milica Janković use lemmas *живот* 'life', *љубав* 'love', *волети* 'to love', *мисао* 'thought' and *осећати* 'to feel' far more often than other authors, and are placed in the same quadrant of the factorial map. On the opposite side of the horizontal axis, in the upper left quadrant, is the lemma *црква* 'church', as well as the author Isidora Sekulić, who uses it more often than other authors. On the other hand, since Isidora Sekulić, apart from the lemma *црква* 'church', very often uses the lemmas *љубав* 'love' and *живот* 'life', as well as the authors Uskoković and Janković, the mentioned authors and the used lemmas are on the same side of the vertical axis. The results of this analysis are gaining their full significance only after an adequate expert interpretation, which goes beyond the scope of this paper.

In the last step, cluster analysis of the matrix obtained by the previously conducted correspondence factor analysis was also performed. The tree diagram (Figure 8) shows the hierarchical grouping based on the relations existing between the author's texts and the lemmas used in them. This classification of texts provides a better understanding and simpler interpretation of the correspondence factor analysis results.

4 Conclusion

This paper presents the current version of the srpELTeC corpus, consisting of Serbian prose literature from the late 19$^{\text{th}}$ and early 20$^{\text{th}}$ centuries. To illustrate the possibilities of the textometric approach, the analysis of the srpELTeC corpus was performed within the TXM programming environment,

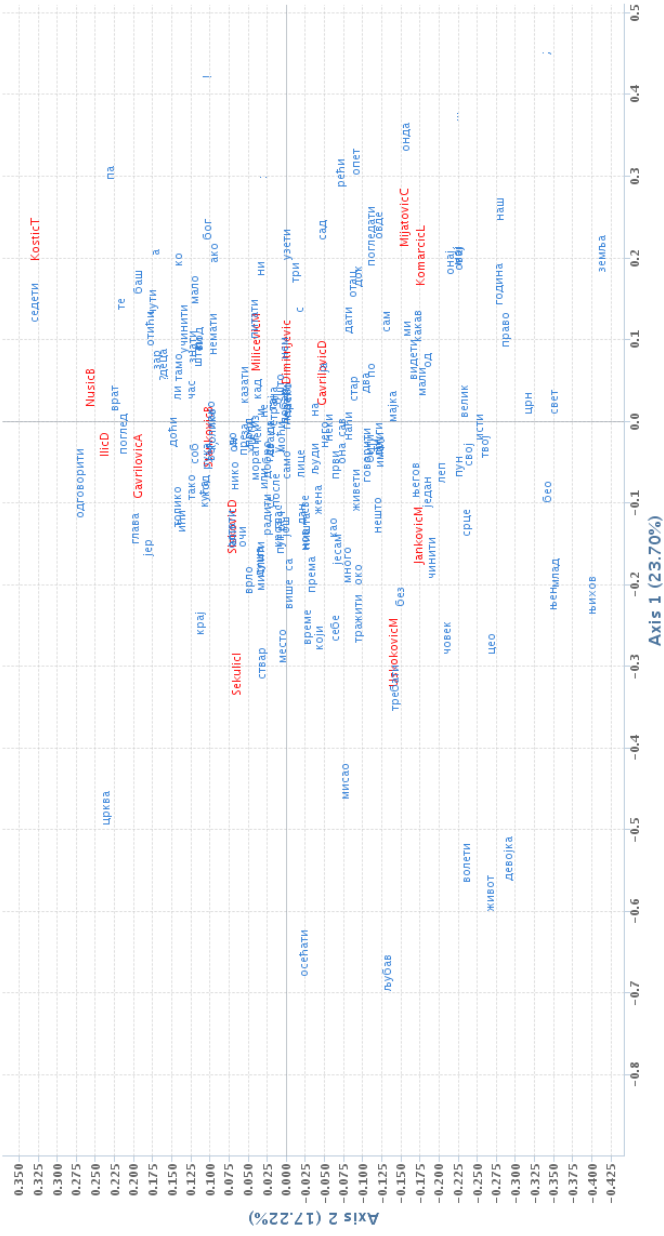

Figure 7. The result of the correspondence factor analysis applied over the corpus author-based parts

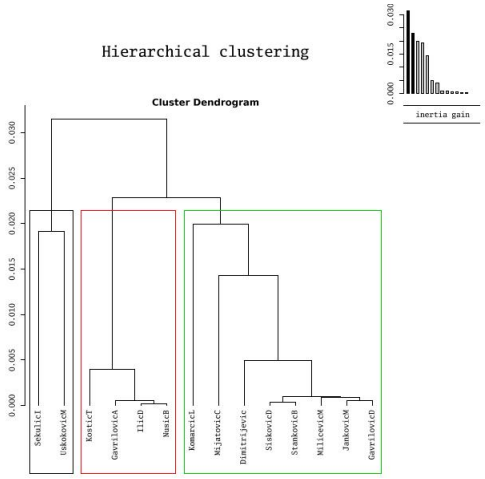

Figure 8. Cluster analysis conducted over the correspondence factor analysis results

presenting the visualization possibilities of the obtained results. The srpELTeC corpus analysis, or some scenarios for the TXM tools application, has no other purpose but to demonstrate the possibilities of using a tool that indicates the significance of textuality and suggests some directions of analysis of those parts that expose and arise from the projections of the corpus itself.

The textometric approach has been used for a long time as a useful method for analyzing corpora of different fields of humanities and social sciences. The laws and conclusions derived from textometric research are based on qualitative and quantitative analysis. The qualitative analysis allows establishing initial hypotheses, which can then be tested on a larger sample by quantitative analysis. The purpose of quantitative or statistical methods is to point out those places in the text that differ and deviate in particular properties. This different way of reading the texts enables asking new questions in the right places, not to give answers, but to identify places that need to be read again and further analyzed, leading to valid interpretation.

Acknowledgment

The author thanks to the COST Action 16204 – *Distant Reading for European Literary History* support, which made possible this research and the author's visit (STSM-CA16204-42562) to the IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités) laboratory, École Normale Supérieure de Lyon, France. The author especially thanks to the hosts Serge Heiden and Bénédicte Pincemin for their hospitality and helpful comments and suggestions.

References

Beaudouin, Valérie. "Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis.". *Glottometrics* Vol. 33 (2016): 56–72

Benzécri, Jean-Paul. *L'analyse des données*. Vol. 2, Dunod Paris, 1973

Dobrić, Nikola. "Korpusna lingvistika kao osnovna paradigma istraživanja jezika". *Naučnostručni časopis za jezik, književnost i kulturu Philologia* Vol. 7 (2009): 47–57

Evert, Stefan and Andrew Hardie. "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium", 2011

Guiraud, Pierre. *Les caractères statistiques du vocabulaire*. Presses universitaires de France, 1954

Guiraud, Pierre. *Problèmes et méthodes de la statistique linguistique*. D. Reidel Publishing Company, 1959

Heiden, Serge. "The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme". In *24th Pacific Asia conference on language, information and computation*, 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010

Heiden, Serge, Jean-Philippe Magué and Bénédicte Pincemin. "TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement". In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, Vol. 2, 1021–1032. Edizioni Universitarie di Lettere Economia Diritto, 2010

Hunston, Susan. *Corpora in applied linguistics*. Ernst Klett Sprachen, 2002

Katnić-Bakaršić, Marina. *Lingvistička stilistika*. Budimpešta: Open Society Institute, 1999

Krstev, Cvetana. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Belgrade: University of Belgrade, Faculty of Philology, 2008

Lafon, Pierre. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots. Les langages du politique* Vol. 1, no. 1 (1980): 127–165

Lafon, Pierre. *Dépouillements et statistiques en lexicométrie*, Vol. 24. Paris: Slatkine-Champion, 1984

Lavrentiev, Alexei, Serge Heiden and Matthieu Decorde. "Analyzing TEI encoded texts with the TXM platform". In *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*, 2013

Lebart, Ludovic and André Salem. *Analyse statistique des données textuelles: questions ouvertes et lexicométrie*. Dunod Paris, 1988

Lebart, Ludovic and André Salem. *Statistique textuelle*. Dunod Paris, 1994

Muller, Charles. *Initiation au méthodes de la statistique linguistique*. Classiques Hachette, 1973

O'Keeffe, Anne and Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010

Pincemin, Bénédicte. "Sept logiciels de textométrie", , 2018, URL https://halshs.archives-ouvertes.fr/halshs-01843695, working paper or preprint

Schmid, H. "TreeTagger-a language independent part-of-speech tagger". *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/* (1994), URL https://ci.nii.ac.jp/naid/20000989946/en/

Utvić, Miloš. "Izgradnja referentnog korpusa savremenog srpskog jezika". PhD. thesis, Univerzitet u Beogradu, Filološki fakultet: Beograd, 2014

Utvić, Miloš. "Annotating the corpus of contemporary Serbian". *INFOtheca* Vol. 12, no. 2 (2011): 39–51