Application of TXM Tools for Spatial Plan Corpus Analysis

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2022.22.1.2

Milena Milinković
milenam@iaus.ac.rs
Institute of Architecture and Urban & Spatial Planning of Serbia
Belgrade, Serbia

ABSTRACT: The paper presents quantitative textometric research of the PP-TXM corpus of spatial plans in the TXM environment. Thanks to the application of developed methods and technologies in the field of corpus linguistics and SrpNER tools for marking named entities, different linguistic and statistical characteristics of the textual part of planning documents have been pointed out. Using the TXM tool during the research, the results were obtained that provided insight into the frequency of occurrence of different types of words and named entities, both in the whole corpus, and their distribution and progression by partitions of this corpus.

KEYWORDS: TXM, textometry, SrpNER, digital corpora, spatial plans.

PAPER SUBMITTED: 1 May 2021
PAPER ACCEPTED: 1 June 2022

1 Introduction

Quantitative approach to the analysis of texts requires observation of their statistical characteristics based on pre-defined criteria of classification and quantification (Bjekić et al. 2012). The classical approach of quantitative analysis of texts frequently required the cooperation of a large number of researchers but also hard and long work (Utvić 2013), which required significant material and time resources (Mehl and Gill 2010). With the advent of computers and the development of modern computer technologies, such analyzes have become easier. Regardless of whether it is a classical or computer-aided approach, any linguistic analysis requires the formation of adequate and representative corpora, so that research conducted on them can be applied to other texts that meet pre-set criteria. In linguistics, the corpus can be described as "a collection of texts that are assumed to be representative of a given language, dialect or other subset of languages, and that will be used for linguistic analysis" (Francis 1975, 15). This kind of definition can be applied to each corpus, regardless of the information carrier. However, with the development of computational linguistics and the formation of electronic corpora, a more acceptable definition is "a collection of authentic machine-readable texts that represent a representative sample of a particular language or linguistic variety" (McEnery, Xiao, and Tono 2006, 5). Electronic corpora are, just like pre-electronic corpora, intended for language research, but unlike them, they are adapted to automatic or semi-automatic processing and analysis.

Corpora can be classified in different ways depending on the scope, the number of languages involved, the purpose, the domain, the media and other characteristics (Popović and Vitas 2003). One of the divisions of the corpus is into general-purpose and specialized corpora, respectively domain corpora The Corpus of the Modern Serbian Language, or *SrpKor* for short, is a general-purpose corpus. Its development began in 1981, and the latest version of *SrpKor2021*, was published in 2021. in the length of over 600 million words. This corpus also includes various domain corpora that have been formed over the years. Domain Corpus of Culinary Arts (Витас and Крстев 2016; Витас 2018; Vitas 2019, 2022), domain corpus of legal texts of the National Assembly of the Republic of Serbia (Васиљевић 2014), domain corpus in the field of librarianship (Трговац 2017) and domain corpus in the field of mining (Обрадовић et al. 2017) are just some of the specialized corpora developed over the years by members of the Society for Language Resources and Technologies - JepTex.[1] The last domain corpus formed within the Society, at the end of 2021, is the domain corpus of spatial planning. Research on the SrpNER and TXM tools was performed on a subset of this corpus, which will be presented in the following chapters.

2 TXM tool

TXM[2] is open source software (Heiden 2010) whose application is enabled in three different operating systems: 64-bit Windows, Mac OS X and 64-bit Linux. There are two versions of this software. The desktop version requires

[1] Society for Language Resources and Technologies – JePTex

[2] What is TXM?

the installation of TXM software, after which the user is allowed to import his corpus on his own computer and perform the desired analysis, while for the TXM version for web software, installation is not required, but the user accesses various online corpora.

If texts are going to be imported into the TXM environment, it is necessary for them to be in a certain format and adequate code layout. With the TXM tool it is possible to analyze documents in TXT plain text format or XML (*eXtensible Markup Language*) format. The first format must be in the recommended UTF-8 code layout, and the second, in addition to the specified code layout, must be in accordance with the TEI (*Text Encoding Initiative*)[3] instructions. The degree of complexity of the analysis, which can be performed in a TXM environment, depends on the level of text representation. Only basic analyzes can be performed on texts in TXT format, since they have the lowest level of representation. In contrast, texts encoded in XML-TEI format can have more or less rich representation, which directly affects the complexity of the analyzes that can be applied to texts.

TXM tool provides users with a variety of techniques for statistical processing of texts, and its working environment is extremely clear and easy to use which makes it suitable for use by both beginners and professionals engaged in statistical research. Since in the TXM environment it is possible to hierarchically organize the text objects on which the research is conducted, the analysis can be performed both on the whole corpus and on a separate subcorp or partition. This is possible because of the fact that corpora are divided into constituent units: textual, structural and lexical units. A textual unit is made of all texts contained in it, marked with the necessary metadata. The structural units of the corpus are chapters, paragraphs, sentences, etc. The lowest level of hierarchical organization consists of lexical units, ie the words of which the text consists (Jaćimović 2019). The results of the given queries are displayed as concordances, tables or graphs.

3 PPTXM corpus

For the purposes of this research, the PPTXM[4] corpus was formed, which consists of texts of spatial plans and is part of the domain corpus in the field of spatial planning. Planning documentation is a specific type of documents characteristic of the field of spatial planning. There are four types of spatial plans that are made on the territory of the state. These are the Spatial

[3] TEI (Text Encoding Initiative).

[4] Corpus of spatial plans formed for analysis in TXM environment.

Plan of the Republic of Serbia, the regional spatial plan, the spatial plan of local self-government units and the spatial plan of the special purpose area (Службени гласник РС 52/2021). Regardless of the plan in question, after making a decision on the development of the plan, it is essential to determine the scope of the plan, make an analysis of the current situation, and then in accordance with the conditions and guidelines specified in higher level planning documents. This whole process requires the cooperation of spatial planners and scientists and experts from many other fields (architecture, urbanism, demography, ecology, sociology, etc.). In the initial stages of drafting the plan, an elaboration for early public insight is prepared in order to acquaint the general public with the purpose and goals of drafting the plan, with potential concepts and solutions for the development of a specific area and the offered environmental plan. This is followed by preparatory activities, which include obtaining the basis for the development of the graphic part of the plan and obtaining the conditions and data necessary for the preparation of the draft planning document containing the textual and graphic part. The draft plan is subject to expert control that lasts 15 days, and after acting in accordance with the report on the performed expert control, the plan is presented to the public at the public inquiry. During this phase of the plan development, interested natural and legal persons, exclusively in writing, submit any objections to the planning document. After the drafting is completed, the plans are adopted by the competent institutions, and then the textual part of the plan is published in the appropriate official gazette (Службени гласник РС 32/2019).

The developed corpus consists of six spatial plans, four plans of the special purpose areas, one regional spatial plan and one spatial plan of the local self-government unit, namely: *Regional Spatial Plan for the Zlatibor and Moravica administrative districts*, *Spatial plan for the special purpose area of Đerdap National Park*, *Spatial plan for the special purpose area of Grlište Reservoir basin*, *Spatial plan for the special purpose area of Ćelije reservoir basins*, *Spatial plan for the special purpose area of international waterway E-80 – Danube (Pan-Europian Corridor VII)* and *Spatial plan for the municipality of Knjaževac*. The only spatial plan that is not included in this corpus is the *Spatial Plan of the Republic of Serbia*, however, bearing in mind that these plans cover different areas in Serbia, and that they cover all the diversity of geographical features of the Republic, it can be said that this corpus representative.

4 Textometry with TXM tool

Textometry is a methodology for analyzing textual data whose development was started by French theorists Pierre Guiraud, Charles Muller, Jean-Paul Benzécri and others, and it is precisely this that has found its application in quantitative linguistic research (MacMurray and Leenhardt 2011), but also in many other humanities and social sciences (Jaćimović 2019). The textometric analysis, which was applied to this corpus, will indicate the different linguistic and statistical characteristics of the textual part of the planning documents. Since this method was founded in the 1980s, thanks to a rich range of tools and software used first for processing and then analyzing natural languages, as well as a number of international standards for structuring data and text, it represents much more than just counting words (Pincemin, Heiden, and Decorde 2020). Different types of textometric analysis are enabled by TXM software. Some of the exemplified textometric analysis in the TXM environment were conducted on one of the domain corpora developed by the Society for Language Resources and Technology. This is the SrpELTeC[5] corpus (Trtovac, Milnović, and Krstev 2021), which at that time contained 21 prose works in the Serbian language published in the period 1840-1920. The motive for its creation was the inclusion of corpus texts in the multilingual collection of European literary texts (*European Literary Text Collection*). The obtained results provided insight into the great variety of analyzes that can be performed with TXM tools, while pointing out the specifics of using different types of words depending on the gender of the author and the frequency of use of certain lemmas in srpELTeC corpus texts (Jaćimović 2019).

5 SrpNER tool and marking of named entities

One of the many tasks of natural language processing is the recognition of Named Entity Recognition (NER). This type of text processing and analysis refers to the recognition of the names of persons, organizations, locations as well as various numerical expressions including percentages, money, time and dates. Recently, the technology of recognizing named entities has been applied to extracting the names of events, products, book titles and marking e-mail addresses. The development of this segment of natural language processing has been going on for almost three decades (Maurel, Friburger,

[5] The complete collection available at SrpELTeC

and Eshkol-Taravella 2014), and the SrpNER system has been developed for many years to recognize named entities in the Serbian language within the Society for Language Resources and Technologies. The manner of work and use of the system for recognizing named entities in the Serbian language was described in detail by prof. Dr. Cvetana Krstev (Krstev et al. 2014).

Prior to any textometric analysis in the PPTXM corpus, the named entities were annotated with the SrpNER tool. For the purposes of this research, geopolitical terms (`top.deogr`, `top.dr`, `top.geo`, `top.gr`, `top.hyd`, `top.reg`, `top.supreg`, `top.ul`), demons (`demonym`) and names of organizations (`org.com`, `org.gen`, `org.pol`, `org.rel`) were tagged. In order for the TXM tool to perform corpus research, certain corrections were necessary, which meant that the dot from the tag for marking the named labels was replaced with an underscore, i.e. the tags were renamed so that the `org.com` tag would be replaced by the `org_com` tag, `top.gr` with `top_gr` label and so on. The only tag that has remained unchanged is the `demonym` tag (Table 1). In addition to the above, it was necessary to eliminate nested tags of named entities, which are produced by the SrpNER tool, and which are not supported in the TXM environment. The annotation of the organization *JKP "Vodovod" Zaječar* with the SrpNER tool was made as follows:

```
<org.com>
  <org.gen>JKP „Vodovod"</org.gen>
  <top.gr>Zaječar</top.gr>
</org.com>
```

Within the external tags used for marking commercial organizations `<org.com>` and `</org.com>` there are also nested tags for general organizations `<org.gen>` and `</org.gen>` which marked *JKP "Vodovod"* and tags for urban settlements or populated places `<top.gr>` and `</top.gr>` which mark *Zaječar*. By removing the nested tags, the full name of the organization is marked with the `<org.com>` tag.

6 Textometric processing and analysis of PPTXM corpora in TXM environment

The first step after importing the corpus into the TXM environment involved automatic segmentation, tokenization, lematization, and finally word type labeling (Jaćimović 2019). This is made possible by the TreeTagger[6]

[6] TreeTagger - a part-of-speech tagger for many languages

TXM tags	Explanation of tags	Example
`demonym`	nouns denoting population, ethnic groups and adjectives derived from geographical names	*beogradski*
`org_com`	commercial organizations	*JKP "Vodovod" Zaječar*
`org_gen`	general organizations (unclassified organizations)	*Institute of Architecture and Urban & Spatial Planning of Serbia*
`org_pol`	political organizations	
`org_rel`	religious buildings (or organizations)	*Church of St. Mark*
`top_deogr`	part of urban settlements or inhabited areas	*Dedinje*
`top_dr`	country	*Serbia*
`top_geo`	geographical features (lowlands, mountains, hills, plateaus...)	*Avala*
`top_gr`	urban settlements or inhabited areas	*Belgrade*
`top_hyd`	Hydronyms (rivers, lakes, springs...)	*Danube*
`top_reg`	region within a state	*Podunavlje*
`top_supreg`	supranational region	*Europe*
`top_ul`	streets or city locations in general	*Knez Mihailova Street*

Table 1. TXM tags with explanations and examples

tool integrated into TXM software that associates each word in the text with the lemma and associated morphosyntactic categories. Such an annotated corpus is suitable for analysis by the *Corpus Query Processor* (CQP), which instead of a character as a search unit includes a corpus word, which provides the opportunity for various phraseological research.

In order to obtain basic corpus structure data within the TXM tool, a search was performed using CQP queries that resulted in a list of concordances. The query for paragraphs `/region[p]` resulted in 11,895 concordances, while the query `/region[s]` related to sentences yielded 17.063 concordance lines.

By tokenizing the texts, it was determined that there are 487,213 tokens in the PPTXM corpus (Table 2). Of that number, there are 32,565 different corpus words, and 14,903 different lemmas. When the punctuation marks are subtracted from the total number of tokens, the corpus consists of 404,218 words. The number of words as well as the number of tokens are not equally distributed among the partitions of this corpus (Figure 1, Table 2).

Name of the spatial plan	**Scope of the plan**	**Dimensions of partitions in number of tokens**
Regional Spatial Plan for the Zlatibor and Moravica administrative district	9.184 km^2	90.724
Spatial plan for the special purpose area of Đerdap National Park	1542 km^2	83.058
Spatial plan for the special purpose area of Grlište Reservoir basin	400 km^2	53.907
Spatial plan for the special purpose area of Ćelije reservoir basins	935 km^2	66.892
Spatial plan for the special purpose area of international waterway E-80 - Danube (Pan-Europian Corridor VII)	4.536 km^2	104.284
Spatial plan for the municipality of Knjaževac	1.202 km^2	88.348

Table 2. Basic geographical and linguistic characteristics of the PPTXM corpus

As already mentioned, the PPTXM corpus consists of six plans (Table 2). The territorial coverage of plans differs significantly from plan to plan, as do the dimensions of texts measured by the number of tokens. The territories covered by the plans range from 935 km2 to 9,184 km2, and the dimensions of the texts range from 53,907 to 104,284 tokens. However, it is noticeable that the size of the territory covered by the plan is not proportional to the dimension of the planning texts, which can be seen in the example *RSP for the Zlatibor and Moravica administrative districts*, which covers 9,184 km2 of territory and whose plan contains 90,724 tokens. Also, *Spatial plan for the special purpose area of international waterway E-80- Danube (Pan-Europian Corridor VII)* whose territory is almost twice as small in area, and the text of the plan contains about 14,000 tokens more. A similar example is with *Spatial plan for the special purpose area of Đerdap National Park* and *Spatial plan for the municipality of Knjaževac.* The first of these plans is larger in terms of territory, and the second contains a larger number of tokens.

The next step in the analysis of the corpus was to determine the frequency of occurrence of punctuation marks and different classes of words in texts. This was achieved through the use of morphological tags (*Part of Speech*

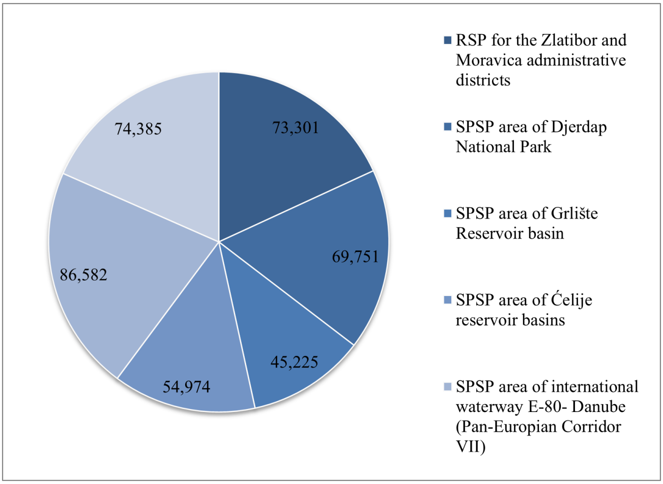


Figure 1. Extracted number of words per partition in the PPTXM corpus.

tags) for the Serbian language. Based on the obtained results (Table 3), it can be seen that the frequency of nouns is the highest, followed by the frequency of punctuation marks, and then the absolute frequencies of other pars of speech are listed. In addition to the total number of nouns, the query `[srpos = "NOUN"]` determined that there are 110,587 different forms, as well as 5,204 different lemmas.

The same query established which nouns are the most common in the entire corpus, and the five most frequent ones are shown in Figure 2, which shows that the frequency of nouns *područje* (area), *zaštita* (protection) and *razvoj* (development) in the corpus is higher than 3,000 and that they are used much more often in plans.

As already mentioned, there are a number of tokens within each partition. In the same way, it is possible to determine how each part of speech is represented in each of the texts of the PPTXM corpus. Considering the ratio of the number of nouns and tokens per corpus partition, it can be concluded that this relationship is disturbed only when it comes to the *RSP of the Zlatibor and Moravica administrative districts* and the *SP of Knjaževac*. According to the number of tokens, the *SPSP of the Zlatibor and Moravica administrative districts* is on the second place, and according to the number of nouns on the third place, while in the case of *SP of Knjaževac*, the situation

srpos	F	srpos	F
N (noun)	153.857	DET (determiner)	7.624
PUNCT (punctuation mark)	82.995	ADV (adverb)	7.367
ADJ (adjective)	78.039	AUX (auxiliary verbs)	6.991
ADP (prepositions)	42.125	PART (particle)	5.917
CCONJ (coordinate conjuction)	33.368	X (other)	3.878
PROPN (proper name)	27.500	SCONJ (subordinate conjunctions)	2.261
VERB (verb)	17.865	PRON (pronoun)	893
NUM (number)	15.873	INTJ (interjection)	660

Table 3. Frequency list of all represented values of the `srpos` position attribute in the PPTXM corpus

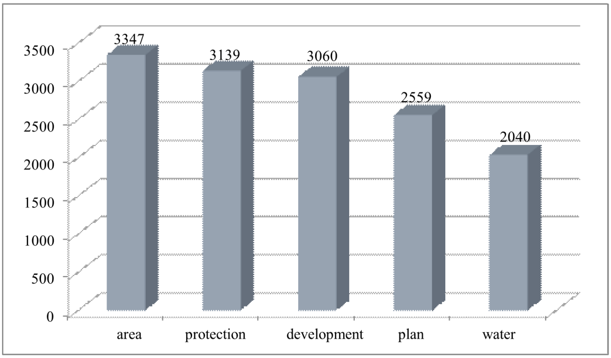


Figure 2. Distribution of nouns whose frequency in the corpus is higher than 2,000.

is the opposite. Other plans have maintained a proportional relationship between tokens and nouns (Figure 3).

The situation is somewhat different when it comes to the relationship between tokens and pronouns in PPTXM corpus partitions. The total number of tokens, as already determined, in the whole corpus is 487.213, and the total number of pronouns obtained by the query `[srpos = "PRON"]` is 893. The representation of tokens by partitions is not proportional to the representation of pronouns. Figure 4 shows the relationship between tokens and pronouns in each of the six spatial plan texts. The highest frequency of use of pronouns is in the *SP for the special purpose area of Ćelije reservoir basins*, although the number of tokens in this plan exceeds only the

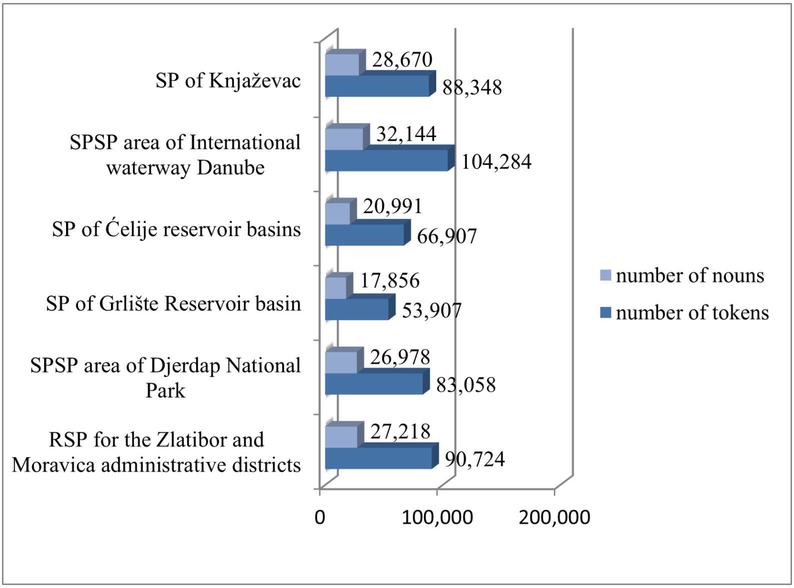


Figure 3. Relationship between tokens and nouns by PPTXM corpus partitions.

number of tokens in the *SP for the special purpose area of Grlište reservoir basin.* On the other hand, the *Spatial plan for the special purpose area of international waterway E-80- Danube (Pan-Europian Corridor VII)* as the largest partition calculated on the basis of the number of tokens, is only in fourth place in terms of the percentage of pronouns. *Spatial plan for the municipality of Knjaževac* and *SP for the special purpose area of international waterway E-80- Danube (Pan-European Corridor VII)* according to the number of tokens, take the third and fourth place, respectively, with a difference of only 5,000 tokens. However, in the *SP for the special purpose area of international waterway E-80- Danube (Pan-European Corridor VII)* the representation of pronouns is almost two and a half times higher than in the *SP for the municipality of Knjaževac.*

Previous analyzes are the basic form of statistical processing of texts. Tools in the TXM environment enable much more complex analysis of various forms of frequency and representation in texts. What is characteristic of the TXM environment is the fact that in the corpus, in addition to the total appearance of a certain part of speech and the frequency of individual forms,

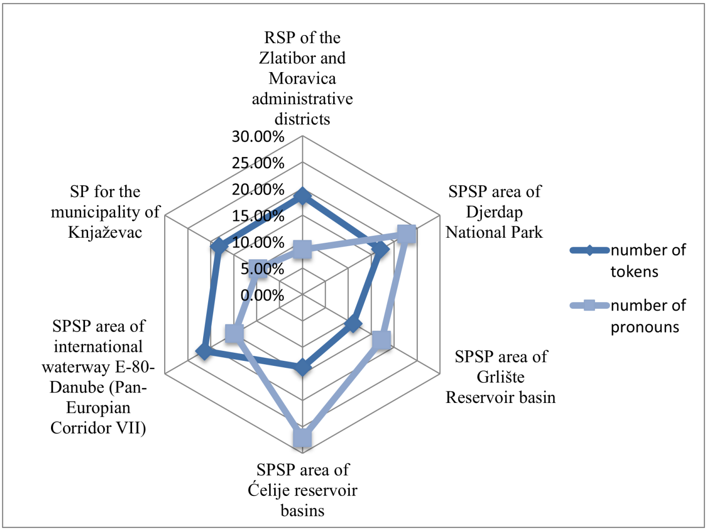


Figure 4. Relationship between tokens and pronouns by PPTXM corpus partitions.

it is possible to monitor the progression and cumulative frequency of different part of speech through corpus components and throughout the corpus.

The graph of progression, shown in Figure 5, gives examples of five nouns: *планина* (mountain), *експлоатација* (exploitation), *водоток* (watercourse), *пристаниште* (port), *споменик* (monument). The red line shows the cumulative frequency of the noun *планина* (mountain). One can notice that it most often occurs in *SP for the municipality of Knjaževac*, while its appearance in *SPSP area of Đerdap National Park* (5) and *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)* (3) is at the level of statistical error. In the remaining three plans, the noun *планина* (mountain) is the least used in the *SPSP area of Ćelije raservoir*, somewhat more often in the *SPSP area of Grlište reservoir basin*, and most often in the *RSP of Zlatibor and Moravica administrative districts*. This frequency of the noun *планина* (mountain) according to the partitions of the PPTXM corps is expected because the area of municipality of Knjaževac, as well as the *RSP of the Zlatibor and Moravica administrative districts* have a pronounced hilly-mountainous relief.

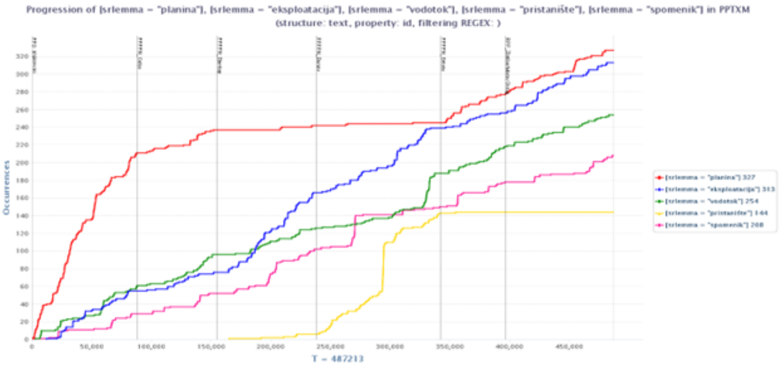


Figure 5. The progression of five nouns in PPTXM corpus texts; from left to right partitions are: *SP for the municipality of Knjaževac*, *SPSP area of Ćelije raservoir*, *SPSP area of Đerdap National Park*, *SPSP area of international waterway E-80 - Danube*, *SPSP area of Grlište*, *RSP of the Zlatibor and Moravica districts.*

The noun *пристаниште* (port) (yellow line) appears 144 times in the whole corpus, however, as can be seen in the graph, this noun was not used in *SP for the municipality of Knjaževac*, *SPSP area of Ćelije reservoir basin and RSP of the Zlatibor and Moravica administrative districts*, barely mentioned in *SPSP area of Grlište reservoir basin*, slightly higher in the *SPSP area of Đerdap National Park*, and the only spatial plan in which the presence of this noun is prominent is the *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)*. Having in mind that this spatial plan covers the entire area of the Danube region in the Republic of Serbia, i.e., the entire course of the Danube in the country, it is completely understandable that the noun *пристаниште* (port) is used throughout the Plan.

The noun *водоток* (watercourse) (green line) is represented in all six plans, and although its appearance in *SP for the municipality of Knjaževac* and *the SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII)* is more prominent, the use of this noun is not negligible in the remaining four plans. Since the entire territory of the Republic of Serbia is characterized by exceptional hydro potential and a developed river system, such a frequency in all six corpus texts could have been expected.

The distribution of natural and cultural heritage on the territory of the state of Serbia can be confirmed by the frequency of the appearance of the noun *споменик* (monument) (purple line) at the level of the entire corpus, but also its constituent parts. The noun *експлоатација* (exploitation) (blue line) is another noun whose presence is noticeable in all PPTXM corpus texts. Of course, this was influenced by the fact that Serbia has certain reserves of different types of mineral raw materials and that their exploitation is present in all parts of the Republic.

The TXM tool provides another important type of analysis that provides the ability to track the distribution of the frequency of occurrence of certain words by partitions. Figure 6 shows the *планина* (mountain), *слив* (basin) и *подручје* (area). Each of them has a different degree of representation in the whole corps. With 3,139 occurrences, the noun *подручје* (area) is the most common noun at the level of the entire corpus, while the noun *слив* (basin) with 620 and *планина* (mountain) with 327 occurrences is much less used in planning documents. The results, after processing in the TXM environment, show a very high frequency of the noun *планина* (mountain) in *SP for the municipality of Knjaževac*, significantly above average compared to its use at the level of the whole corpus, while its frequency is below average in the remaining five spatial plans, especially in *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII).* The use of the noun *слив* (basin) is most pronounced in the *SPSP area of Ćelije reservoir basin* and there is a positive score in the *SPSP area of Grlište reservoir basin.* In all four remaining plans, the use of this noun is below average, and as with the noun *планина* (mountain), it is most prominent in the *SPSP area of international waterway E-80 - Danube (Pan-European Corridor VII).* The noun *подручје* (area) has a negative representation, when comparing with the entire corpus, in the *RSP of the Zlatibor and Moravica administrative districts* and the *SP for the municipality of Knjaževac.* In other partitions, the representation of this noun is above average.

Bearing in mind the fact that the TXM platform imported a corpus previously marked with named entity tags, it can also analyse, in addition to determining the frequency of different part of speech, the degree of representation of named entities, using the `Index` option and posing different queries, such as `/region[top_gr]`, `/region[top_hyd]`, `/region[demonym]` and others. Figure 7 shows the six most frequent named entities in the PP-TXM corpus. In addition to the total frequency, the numbers of different lemmas and different forms for each named entity are shown. By far the most represented named entity type is `top_gr`, which is used for urban settlements and other inhabited areas.

The query `/region[org_rel]` determined that there are 33 different forms of named entities `org_rel`, which marked religious objects, and that 29 different lemmas with a total of 56 occurrences were recorded for them.

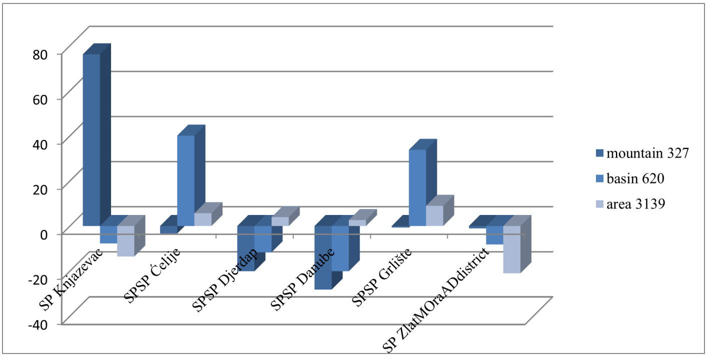


Figure 6. : Specificity of the use of nouns in the PPTXM corpus.

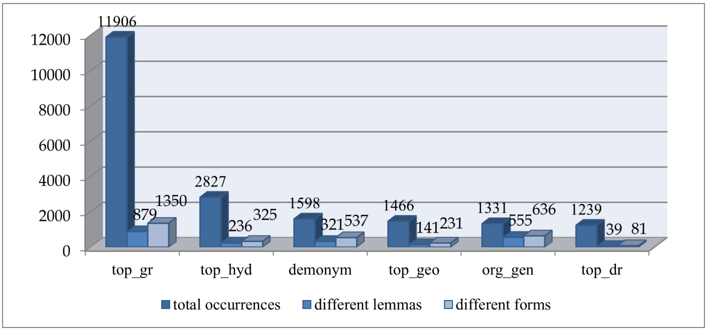
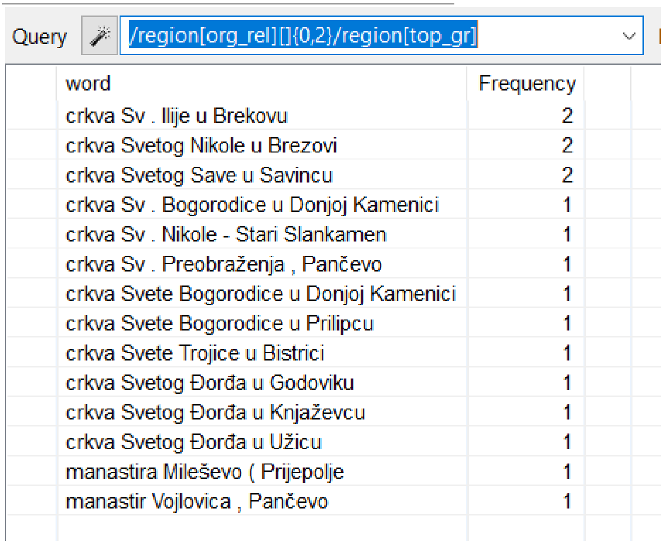


Figure 7. Most represented named entities in the PPTXM corpus.

The most frequent are *Манастир Милешева* (Mileseva Monastery) (6), *Црква Св. Петра* (the Church of St. Peter) (4), *Црква Св. Николе* (Church of St. Nikola), *Св. Тројице* (St. Trinity), *Св. Ђорђа* (St. George) and *Манастир Рача* (Monastery of Rača) with 3 appearances, then eight religious buildings that appear in the corpus 2 times and the remaining named entities `org_rel` that appear only once. However, in the TXM environment, it is possible to formulate more complex queries. Thus, the query `/region[org_rel][]0,2/region[top_gr]` determines the frequency of named entities `org_rel` followed by the named entity `top_gr`. The result is a completely different order in the number of appearances of religious buildings, as shown in Figure 8. In this case, the most frequently listed are *Црква Св. Илије* (the Church of St. Elijah) in Brekovo, *Црква Светог*

Николе (the Church of St. Nicholas) in Brezova and *Црква Светог Саве* (the Church of St. Sava) in Savinac.



Query /region[org_rel][]{0,2}/region[top_gr]

word	Frequency
crkva Sv . Ilije u Brekovu	2
crkva Svetog Nikole u Brezovi	2
crkva Svetog Save u Savincu	2
crkva Sv . Bogorodice u Donjoj Kamenici	1
crkva Sv . Nikole - Stari Slankamen	1
crkva Sv . Preobraženja , Pančevo	1
crkva Svete Bogorodice u Donjoj Kamenici	1
crkva Svete Bogorodice u Prilipcu	1
crkva Svete Trojice u Bistrici	1
crkva Svetog Đorđa u Godoviku	1
crkva Svetog Đorđa u Knjaževcu	1
crkva Svetog Đorđa u Užicu	1
manastira Mileševo (Prijepolje	1
manastir Vojlovica , Pančevo	1

Figure 8. Frequency of religious objects based on query `/region[org_rel][]0,2/region[top_gr]`.

7 Conclusion

The paper presents various linguistic and statistical analyzes, using TXM tools, on the texts of spatial plans in which the named entities were previously marked using SrpNER, a system for recognizing named entities in the Serbian language. The common characteristics of the texts of planning documentation were pointed out, as well as certain particularities of individual spatial plans, which are included in the Serbian corpus *SrpKor2021*, which currently contains more than 600 million words, as part of the domain corpus of spatial planning.

Spatial plans, which are the main focus of this paper, are a specific type of documents characteristic in the field of spatial planning. In order to enable more detailed research of these texts, the PPTXM corpus was created, which includes six documents dealing with different regions of the Republic of Serbia: Zlatibor and Moravica administrative districts, basins of reservoirs “Grlište” and “Ćelije”, Đerdap National Park, International Waterway E-80 - Danube (Pan-European Corridor VII) and the municipality of Knjaževac.

The analysis covered more than 400,000 words, whose representation ranges from 11.18% in the *PPPPN of the “Grlište” reservoir basin* to 21.42% in the *PPPPN of the international waterway E-80 - Danube (Pan-European Corridor VII).* The corpus contains more than half a million tokens, a third of which are nouns.

It should be noted that, in addition to the presented analyzes, the TXM environment provides other possibilities and that their use as well as the use of presented textometric approaches depends on pre-set requirements as well as the needs and preferences of researchers. What is indisputable is that the TXM platform enables the display of the features of the corpus as a whole, as well as its components. For that reason, the application of these methods is very important for various quantitative analyzes and determination of characteristic features of texts. Some of the presented results, such as the representation of specific nouns in the entire corpus and the frequency of their occurrence in individual plans, could be deduced in advance on the basis of the plan coverage area. However, the obtained results on the representation of pronouns by corpus partitions cannot be given a logical explanation only on the basis of knowledge about the scope of the plans; for their better understanding it is necessary to conduct additional research and analysis.

Acknowledgment

This research was conducted in the scope of the work on a doctoral dissertation at the Faculty of Philology, University of Belgrade entitled “The Development of library and language resources for organizing and finding information on spatial planning.” The tools used during the research were specifically developed for the Serbian language (SrpNER) or were adapted for various linguistic research of Serbian texts (TXM) within the Society for Language Resources and Technologies. The author owes special gratitude to professors Dr. Cvetana Krstev and Dr. Ranka Stanković, without whose personal engagement, the research presented in this paper would not have been possible.

References

Bjekić, Jovana, Ljiljana Lazarević, Milica Erić, Elena Stojimirović, and Teodora Đokić. 2012. "Razvoj srpske verzije rečnika za automatsku analizu teksta (LIWCser)." *Psihološka istraživanja* XV (1): 85–110. https://doi.org/10.5937/PsIstra1201085B.

Francis, W Nelson. 1975. "Problems of Assembling, Describing, and Computerizing Corpora." *Research Techniques and Prospects. Papers in Southwest English,(1),* 15–38.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation,* 389–398. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://aclanthology.org/Y10-1044.

Jaćimović, Jelena. 2019. "Textometric methods and the TXM platform for corpus analysis and visual presentation." *Infotheca - Journal for Digital Humanities* 19 (1): 30–54. https://doi.org/10.18485/infotheca.2019.19.1.2.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489. https://doi.org/10.1093/logcom/exs079.

MacMurray, Erin, and Marguerite. Leenhardt. 2011. "Textometry and Information Discovery: A New Approach to Mining Textual Data on the Web." In *Proceedings of the 2011 International Conferece on Artifical Intelligence, ICAI 2011,* edited by H.R. Arabnia, D. Fuente, E.B. Kozerenko, and J.O. Olivas, II:605–611. Las Vegas: Worldcomp'11.

Maurel, Denis, Nathalie Friburger, and Iris Eshkol-Taravella. 2014. "Enrichment of Renaissance Texts with Proper Names." *Infotheca* 15 (1): 29a–41a. https://infoteka.bg.ac.rs/index.php/en/archives/2014/1/infoteka-15-1-2014-30-41.

McEnery, Tony, Richar Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book.* London: Routledge. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm.

Mehl, Matthias R., and Alastair J. Gill. 2010. “Automatic text analysis.” In *Advanced methods for conducting online behavioral research,* edited by S. D. Gosling and J. A. Johnson, 109–127. Washington: American Psychological Association. https://doi.org/10.1037/12076-008.

Pincemin, Bénédicte, Serge Heiden, and Matthieu Decorde. 2020. “Textometry on Audiovisual Corpora Experiments with TXM software.” In *JADT 2020 : 15es Journées internationales d’Analyse statistique des Données Textuelles.* http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/PINCEMIN_HEIDEN_DECORDE_JADT2020.pdf.

Popović, Ljubomir, and Duško Vitas. 2003. “Konspekt za izgradnju referentnog korpusa srpskog standardnog jezika.” In *Naučni sastanak slavista u Vukove dane.* Beograd, Novi Sad: MSC.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. “The Serbian Part of the ELTeC - from the Empty List to the 100 Novels Collection.” *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. https://doi.org/10.18485/infotheca.2021.21.2.1.

Utvić, Miloš. 2013. “Izgradnja referentnog korpusa savremenog srpskog jezika.” PhD diss., Univerzitet u Beogradu, Filološki fakulte. https://nardus.mpn.gov.rs/handle/123456789/4091.

Vitas, Duško. 2019. “Food as Text.” *Infotheca - Journal For Digital Humanities* 19 (2): 139–161. https://doi.org/10.18485/infotheca.2019.19.2.7.

Vitas, Duško. 2022. “From Onions to Champagne – Food and Drink in the SrpELTeC Corpus.” *Infotheca - Journal for Digital Humanities* 21 (2): 88–118. https://doi.org/10.18485/infotheca.2021.21.2.5.

Васиљевић, Небојша М. 2014. “Аутоматска обрада правних текстова на српском језику.” PhD diss., Универзитет у Београду, Филолошки факултет. https://nardus.mpn.gov.rs/handle/123456789/4091.

Витас, Душко, and Цветана Крстев. 2016. “Оглед из гастрономатике [Experiments in gastronomatics].” In *Теме језикословне у србистици кроз дијахронију и синхронију [Linguistic topics in Serbian through diachrony and synchrony],* edited by Јасмина Дражић, Исидора Бјелаковић, and Дејан Средојевић, 1–10. Novi Sad: Филозофски факултет.

Витас, Душко M. 2018. “Храна из нежељене поште : (анатомија језика брзе хране [Spam Food: (Anatomy of Fast Food Language)]).” In *Српски језик и његови ресурси: теорија, опис и примене Научни састанак слависта у Вукове дане,* edited by Божо Ћорић and Александар Милановић, 21–35. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. https://doi.org/10.18485/msc.2018.47.3.ch2.

Обрадовић, Иван, Александра Томашевић, Ранка Станковић, and Лазић Биљана. 2017. “Увођење доменских и семантичких маркера за област рударства у српске електронске речнике.” In *Научни састанак слависта у Вукове дане - Српски језик и његови ресурси: теорија, опис и примене,* edited by Рајна Драгићевић and Александар Милановић, 147–158. Београд: Међународни славистички центар на Филолошком факултету. https://doi.org/10.18485/msc.2017.46.3.ch10.

Службени гласник РС. 52/2021. *Закон о планирању и изградњи.* Бр, 72/2009, 81/2009 - испр., 64/2010 - одлука УС, 24/2011, 121/2012, 42/2013 - одлука УС, 50/2013 - одлука УС, 98/2013 - одлука УС, 132/2014, 145/2014, 83/2018, 31/2019, 37/2019 - др. закон, 9/2020 и 52/2021), 52/2021.

Службени гласник РС. 32/2019. *Правилнику о садржини, начину и поступку израде докумената просторног и урбанистичког планирања.* Бр. 32/2019), 32/2019.

Тртовац, Александра. 2017. “Аутоматизација библиотека у Србији - историјски преглед.” *Библиотекар,* no. 2, 101–114. https://bds.rs/wp-content/uploads/2018/01/bibliotekar-2-2017-7.pdf.