Computer Science and Information Systems 15(2):369–392 https://doi.org/10.2298/CSIS170701005P

Use of Linguistic Forms Mining in the Link Analysis of Legal Documents

Đorđe Petrović and Milena Stanković

Faculty of Electronic Engineering, University of Niš,
Aleksandra Medvedeva 14, 18000 Niš, Serbia
petrovicdj@gmail.com, Milena.Stankovic@elfak.ni.ac.rs

Abstract. This document employs a statistical approach in exploring language and extracting linguistic forms there contained, so as to identify the linguistic forms which are most frequently used in legal documents. Thus retrieved data, as the second part of this paper shows, can be used to research information, analyze references and links, trace pathways between correlating legal documents and establish the relevance of legal documents on the grounds of their mutual correlation. The retrieved data can further be utilized in various other manners. The methodology of this research and thus attained information form a good basis and act as input data for numerous further analyses.

Keywords: linguistic forms mining, legal documents mining, text mining, information extraction, natural language processing, link analysis.

1. Introduction

Legislative information is often kept in textual form of relatively unstructured form [27]. For instance, statutes, rulings and explications are commonly stored as free text documents. Such documents can be structured in a variety of ways. The national legislation of the Republic of Serbia contains a document entitled “Unique Instructions on the Methodology for the Creation of Regulations” [24], passed by the Legislation Committee of the National Assembly of the Republic of Serbia. This document prescribes instructions for the creation of laws, decisions, rules of procedure and similar general acts passed by the National Assembly of the Republic of Serbia. Even though it covers all relevant legal aspects of miscellaneous affairs, the instruction does not define the automatic management of complex information contained in such legal documents. Moreover, the national legislation of the Republic of Serbia includes laws and other regulations which had been passed prior to the establishment of the Unique Instructions on the Methodology for the Creation of Regulations [24], which had, therefore, not necessarily been compiled in compliance with these regulations. Consequently, these laws and regulations are diverse both in terms of their structure and content.

In an effort to come to terms with such occurrences, an initiative has been brought forth to standardize metadata of legal documents [32]. The proponent suggests an advanced set of metadata tags and pre-defined legal terms to enable an automatic exchange and linking of legal documents in the Internet to cater for the demands of the wider community. However, until such standards become widely accepted and applied, the legal documents currently in use, which are not structured in such a manner, can still

be explored through the existing search and text analysis tools. The exploration of knowledge through automatic analysis of free text is a field of research which originates from the information retrieval research method and is commonly dubbed *text mining* [27]. Automatic retrieval of information from a large collection of documents was one of the first applications of computer science on legislation, but it remains an important objective today [27]. The two main goals regarding data in this area are [30]:

- improving techniques and methods of managing complex findings in this field, and
- establishing suitable manners of preserving and retrieving data

To a certain extent, the methodology described herein aims to deal with and contribute to resolving both of these issues. It applies certain data science techniques and algorithms in order to extract information from legal texts and other regulations, in an attempt to promote a practical application of automatic management of complex information contained therein. Thus retrieved data can subsequently be utilized to explore information contained in legal documents, analyze references and links within legal documents, trace pathways between correlating legal documents, establish the relevance of legal documents on the grounds of their mutual correlation, as well as many other analyses. The results of this research can especially be helpful when it is necessary to amend laws and other legislation. Such amendments affect all relating regulations and areas of interest. While establishing a system of links between legal documents, laws and other regulations, it is possible to achieve efficient management of amendments to these documents while taking into account their mutual correlations. This is the answer to the first issue involved with the automatic retrieval of information from large document collections [30], which is to improve techniques and methods applied to process complex findings in this field. During its evolution cycle, a law may get modified several times, sometimes dozens of times [8]. Furthermore, with regards to the techniques and methods of managing complex legal documents, the procedure which we have used to transform, or rather restructure data, contains a description of storage and preservation of legal documents. Bearing in mind the fact that the main classification unit of any law is an article, we have segmented law texts into individual law articles, so that each article is preserved as a single unit of record in the data base. We have concluded that this is a suitable method of storing and preserving data from laws and other regulations, as it facilitates further analysis and management of complex legal documents. In this way, we have also addressed the second issue related to the automatic retrieval of information from large collections of documents [30], which is to establish suitable manners of retrieving and preserving data.

The information attained through research of linguistic forms in legal documents is an excellent basis and acts as input data for numerous further analyses.

2. Related Work

There are several studies dealing with research of “linguistic forms”, phrases, “sequences of words” etc. In their book, Liu, et al. [14] investigated automatic identification of high-quality phrases from innumerable documents. In their work, they suggest principles and methodologies to acquire results based on phrases of variable length and present applications which conduct phrase mining.

In addition, there are several studies exploring the frequency of occurrence of linguistic forms in documents, identification of key linguistic forms, identification of the "true quality" of linguistic forms and so forth. In their papers, Karanikolas & Skourlas [11,12] conduct research on automatic classification of documents. They focus on the issue of extracting key-phrases from a collection of texts in order to use them as attributes for text classification. They look for sequences of words (key-phrases) that will be used as features for classification rules and not for extracting association rules. In their works Karanikolas & Skourlas [11,12] extracted key-phrases which are frequent within the documents of one or few classes but are not so frequent in the documents of the remaining classes of the training set. Furthermore, in [12] it is said that words that constitute key phrases must coexist in a specific window size. The authors introduce the authority list creation algorithm (ALCA) that reduces the search space by building larger key phrases from smaller ones. In addition, Liu, et al. [14] state that the raw frequency of data tends to produce misleading quality assessment. The authors attempt to rectify the decisive raw frequency to establish the true quality of a phrase by examining the context of its citations.

Several papers deal with the analysis of legal document texts and the analysis of legal document references. In addition, several papers deal with national legal documents of certain countries. In the work of Saravanan, et al. [26] the authors propose a mechanical learning approach in order to identify rhetorical roles in legal documents. This is accomplished through the extraction and modified ranking of sentences containing additional meanings of their actual roles within a text. They also delved into linguistic forms often used to signal common rhetorical roles in texts.

In the words of Lodder & Oskamp [15], legal documents are filled with implicit and explicit references. There are numerous works describing methods of analyzing links in legal documents. In their work, Opijnen, et al. [20] describe a software framework to detect and pursuit references in national and EU legislations, judicial practice, parliamentary documents and official journals. Monroy, et al. [18] describe the application of algorithms to analyze links (Page Rank) in search of relevant information in legal documents. Furthermore, Neale [19] presents an analysis of quotes from the Canadian case law practice, where he analyzes existing quotes using a web-based system he created.

There are several other studies dealing with legal documents link analysis, such as [3, 25, 31]. In principle, these studies conduct link analysis based on data in XML format which have tags used to denote links between documents. In their paper, Bommarito II & Katz [3] investigate the properties of the United States Code (source of Federal Law) citation network by examining the directed degree distributions of the network. For this research, the authors obtained an XML snapshot of the Code, where citations are explicitly coded within these XML documents at the section level. Sakhaee, et al. [25] studied legislation networks applied in New Zealand in their research. The authors used the XML format from the NZ Government Legislation website, with links between documents. Waltl, et al. [31] used publically available data from German legislation. They analyzed laws with regards to the occurrence of different reference types in legal documents.

In their paper, Koniaris, et al. [13] present an approach for extracting a machine readable semantic representation of legislation, from unstructured document formats. Their method expresses legal documents structure in the form of a set of syntactic rules, i.e., a domain-specific language for legal documents, and their approach was based on a

set of Greek legal documents. The method described in their paper involves a conversion of all legal documents into plain text files. All legacy and plain legal documents are then converted into a standard XML format thus producing Akoma Ntoso compliant XML files. Akoma Ntoso is an XML schema for modeling parliamentary, legislative, and judiciary documents [2]. Our research has not dealt with the internal division of regulations into information blocks or wider classification units. Instead, regulations were divided into articles of law. An article of law is a single logical entity and we conducted a segmentation of law texts into individual law articles.

In their work, Garofalakis, et al. [8] address the issue of different versions of the same laws existing and being stored in the revision control system. They try to answer the question of how to allow users to visit the current version of each law, browse its revision history and track changes between different revisions, so that they could automatically apply modifications and allow their publication in a revision control system.

There are also studies exploring legal documents in the Serbian language. The subject of research presented in 28 concerns specific language rules in legislative texts in the Serbian language that can be expressed by computational linguistic methods. This research meticulously describes referencing phrases and the structure of legal acts by using natural language processing methods based on language rules.

3. Applied Methodology of Legal Documents Analysis

According to [17], there are two strategies for entity extraction from text: Rule-based approach and Statistical approach. Our research, unlike the research presented in [28], makes use of the statistical approach in processing natural language. The same method was also used by Furlan, et al. [7].

Unlike [26], we have not dealt with complete sentences and their roles in texts, but rather explored all linguistic forms found in legal documents.

In contrast to studies dealing with legal documents link analysis, such as [3, 25, 31], our source of information was constituted of textual documents which had no pre-prepared citations and links. We then used this set of documents to exemplify the procedure of identifying citations and links within texts. Apart from presenting the distribution of reliability in-degrees and out-degrees, our research has not delved into further analysis of this distribution.

Our paper has not examined different versions of laws as described by Garofalakis, et al. [8], but explored links between laws to exemplify possible effects that amendments to a certain law can have on a set of other valid laws related with it.

The techniques commonly used for information retrieval and text mining in legal documents and databases are as follows [27]: information extraction, text summarization, text categorization and text clustering. Based on this classification, the methodology herein described can be identified as *"information extraction",* which involves retrieving data from a collection of documents and analyzing existing relations [17]. The main characteristic of information extraction is the completion of a pattern inherent to a certain domain [29]. In our case, it is the legal domain.

Our experiment has followed the following steps or phases:

- Data collection

- Data preprocessing
- Data transformation
- Linguistic forms mining
- Link analysis in legal documents based on retrieved linguistic forms.

3.1. Data Collection

The category of legal documents includes laws, regulations, rule books, decisions, statutes, sentences, orders, instructions and other documents from this domain. This paper and its experiment are limited only to the legislative documents applicable in the Republic of Serbia. Article 29 of the Serbian Law on Publication of Laws and Other Regulations and Acts, states that all Internet users are granted free access to all regulations and other legal acts valid in the Republic of Serbia. Consequently, all bills of law are made publically available in several public web sites, including the web sites of The National Assembly of the Republic of Serbia (http://www.parlament.gov.rs/akti/doneti-zakoni/doneti-zakoni.1033.html), The Legal Information System of the Republic of Serbia (http://www.pravno-informacioni-sistem.rs/SlGlasnikPortal/reg/content) and others. As these sites contain complete bills of law, we gathered a collection of 1,120 texts of valid laws for the purpose of this paper. These laws were gathered in 2016, at which time they were a collection of all valid laws. From that moment on, the number of valid laws has changed, as certain new laws have been adopted, some have been altered, while others have ceased to be valid. Our research was conducted on the collection of laws gathered in the described manner and has not taken into account any subsequent alterations.

Even though the above-mentioned web sites have laws divided into various categories and subcategories, this aspect was not taken into consideration when compiling this paper.

3.2. Data Preprocessing

The laws which we gathered had the form of text documents. Most procedures for text preprocessing utilize the method of normalization [17]. Text normalization in the Serbian language, due to the specific nature of the language [29], proves to be a significant challenge [10], and mostly involves the application of algorithms to exclude words which bear little or no context significance, commonly called *stop words*, and algorithms to reduce different forms of individual words to their common root, also known as *stemming*. The use of some of these algorithms is described in studies such as [7] or [16], we have used neither of these algorithms.

The language used in laws and regulations is predominantly formal and comes down to the recurrent usage of similar linguistic forms, which calls for the need to process language expressions and phrases, rather than individual words. According to Feldman & Sanger [5], “phrases, multiword forms, or even multiword hyphenates would not constitute single word-level features”. Therefore, this research has not treated individual words, but rather expressions containing multiple words or multiword forms, or in other words, frequently occurring sentence parts of the same form.

At this stage, the texts were prepared for further analysis by being converted into “plain text”, which disregarded any formatting or text structuring. We then applied a case normalization algorithm, which converts the whole text into lower-case letters, in order to avoid disambiguation of the same text by disregarding different variants of a text written in combinations of upper and lower-case letters [17].

English speakers are familiar with the concept of case. Each letter in the Latin alphabet comes in two "cases": upper and lower (or capital and small). Similarly, the Greek, Cyrillic, and Armenian alphabet are cased, or "bicameral". Occasionally, situations arise where a letter (or a whole string of letters) needs to be converted from one case to the other. In order to erase case distinctions, everything can be converted to upper-case or lower-case. In Unicode, this is done with a different process, known as case folding [9]. There are examples in which converting to any one case will preserve some distinction that converting the other way will erase, and doing it in different orders will produce different results. For example, certain letters of the Turkish alphabet prove difficult to manage in case folding. Similarly to the study of Milošević [16], where text normalization makes use of text conversion into small (lower-case) print, we have used the same procedure. This conversion, as the remainder of our text shows, has not caused any problems in our research.

3.3. Data Transformation

Stranieri & Zeleznikow [27] state that one of the five following methods can be used to transform data: aggregation of data values, normalization of data values, feature reduction, example reduction and restructuring. We find that the segmentation of texts into obviously related sections can be useful from the point of view of data identification and its subsequent referencing. In addition, selecting adequate text sections is useful in cases where either documents are very long, or only parts of documents are of interest to users [22].

Standard internal division of laws and other regulations includes the following wider classification units [24]:

- a part,
- a chapter,
- a section and
- a sub-section.

A part incorporates the thematic entirety of a regulation and represents the broadest classification unit of a regulation. It can be divided into chapters, which are used to separate parts into functional or thematic entireties. Chapters are further divided into sections and sub-sections.

An article of law is a single logical entity containing one or more legal norms. It can be further divided into paragraphs, paragraphs into points, points into sub-points, and sub-points into indents. Furthermore, an article is also the main classification unit of laws. The transformation we applied was aimed at storing an article of law, being a basic classification and logical entity, as a single record into our database. We conducted a segmentation of law texts into individual law articles, thus acquiring a total of 59,046 records in the database. Each of the records is kept alongside information on the name of the law the article belongs to, as well the respective ID of the article.

During our research, we have not dealt with the internal division of regulations into broader classification units, such as parts, chapters, sections and sub-sections. This aspect can certainly be the subject of a different research.

According to the division presented in [5], we used the previously described manner to attain entities, or basic text units, which all laws are comprised of. We will attempt to show further on whether it is possible to distinguish any additional basic elements when exploring linguistic forms of legal documents, in accordance with the previously mentioned division.

3.4. Linguistic Forms Mining

According to the definition of Oxford University Press [21], linguistic form is:

1. A unit or pattern of language, typically observed independently from its associated function or value.
2. The characteristics or form of such units or patterns of language. Also: these characteristics considered in relation to one such unit or pattern of language.

Relevant literature also recognizes other terms bearing the same or a similar meaning, such as "key-phrases", "sequences of words", "word phrases" etc. Our conclusion was that the term "phrases" was certainly not adequate to describe legal terms and expressions being examined in this research. In accordance with its dictionary definition [21], this paper has used the term "linguistic form", adding that the term "sequence of words" is also adequate. In the context of the paper we state that a linguistic form is a frequent word (stem) or sequence of words (key-phrase) where words (stems) are presented in a specific order.

Further extraction from law articles locates and identifies linguistic forms which can also be considered new entities. According to Miner, et al. [17], there are two strategies for entity extraction:

- Rule-based approach, which defines conditional rules to be applied in a text in order to identify possible entities;
- Statistical approach, which treats entity extraction as the process of sequence classification.

Our paper has used statistical approach to linguistic forms extraction, aiming to locate linguistic forms most frequently used in the observed texts. To that extent, we have used the Python programming language to create a program to detect all linguistic forms (sentence parts) in a previously described database. One of the program's parameters is N – input variable containing the number of words a single linguistic form is consisted of. In brief, the following algorithm is applied:

```
for n inrange(2, N):
  for record in databasecursor:
    find all n-word linguistic_forms in current record
    insert into table all n-word linguistic_forms
select and group n-word linguistic_forms, count of n-word
linguistic_forms
```

According to Liu, et al. [14], in practice, one can also set a maximum input variable N to restrict the number of words of a single linguistic form (phrase length). Even if no explicit restriction is added, phrase length is typically a small constant. Firstly, the

program is run to detect all two-word linguistic forms, followed by three-word forms, and so forth. In the course of our research, we detected linguistic forms consisting of up to five words, using N=5 as the input parameter of the abovementioned algorithm. The results are presented in a table within the database containing linguistic forms found in all the articles of observed laws, as well as the number of instances in which an article of a law mentions a linguistic form. Table 1 shows the number of different linguistic forms found in the observed database:

Table 1. Total number of different linguistic forms detected in the observed database

	Max number of linguistic forms
2-word linguistic forms	1,651,855
3-word linguistic forms	3,702,636
4-word linguistic forms	5,021,611
5-word linguistic forms	5,754,751

For further evaluation, the detected linguistic forms are grouped and counted with regards to their occurrence in different law articles. Based on this information, it is now possible to acquire exact data on the extent at which certain linguistic forms are used to design the texts of the observed laws. Table 2 shows some of the most frequently used linguistic forms (English translation and original in Serbian) from the observed database and the respective percentage of their occurrence in law articles or table recordings:

Table 2. The most frequently used linguistic forms (English translation and original in Serbian) in the observed database and the respective percentage of their occurrence in law article

2-word linguistic forms	**3-word linguistic forms**	**4-word linguistic forms**	**5-word linguistic forms**
"this article" ("ovog člana") 26.67%	"in compliance with" ("u skladu sa") 23.66%	"paragraph 1 of this article" ("stava 1 ovog člana") 17.10%	"from paragraph 1 of this article" ("iz stava 1 ovog člana") 15.63%
"in compliance" ("u skladu") 24.58%	"1 of this article" ("1 ovog člana") 18.45%	"from paragraph 1 of this" ("iz stava 1 ovog") 15.65%	"from paragraph 2 of this article" ("iz stava 2 ovog člana") 4.19%
"in compliance with" ("skladu sa") 23.66%	"paragraph 1 of this" ("stava 1 ovog") 17.13%	"in compliance with law" ("u skladu sa zakonom") 6.10%	"in compliance with this law" ("u skladu sa ovim zakonom") 3.79%
"this law" ("ovog zakona") 22.45%	"from paragraph 1" ("iz stava 1") 16.95%	"in compliance with this" ("u skladu sa ovim") 4.68%	"this law coming into action" ("stupanja na snagu ovog zakona") 3.41%

2-word linguistic forms	3-word linguistic forms	4-word linguistic forms	5-word linguistic forms
"from paragraph" ("iz stava") 21.67%	"within (period of time)" ("u roku od") 9.39%	"paragraph 2 of this article" ("stava 2 ovog člana") 4.67%	"date of this law coming into action" ("dana stupanja na snagu ovog") 2.51%
…	…	…	…

4. Legal Documents Link Analysis based on Detected Linguistic Forms

Our further analysis will attempt to exemplify the significance and possible applications of the retrieved data. In the words of Lodder & Oskamp [15], legal documents are filled with implicit and explicit references, which is a statement confirmed through our research.

The extracted linguistic forms were stored in a relevant database. It is immediately noticeable that the most commonly found forms are used for referencing or linking to the same or other sub-articles, articles or laws. Within the compilation of most frequently occurring linguistic forms, we performed a manual data mining in order to detect linguistic forms of similar content, but showing a lower frequency of occurrence in the given collection of detected linguistic forms. We applied SQL queries. An example of an SQL query applied to detect 3-word linguistic forms used to establish links with other laws, which are at the same time similar to expressions such as "in compliance" (the most frequently occurring 2-word linguistic form in the Serbian language, used to establish links with other laws) is given below:

```
SELECT * FROM tblExtractedLF
WHERE ((tblExtractedLF.LinguisticForms)
Like "*in compliance*");
```

This way, we were able to identify a number of 3-word expressions used for linking with other laws, such as: "in compliance with" (which is also the most frequently occurring 3-word linguistic form), "compliance with law", "compliance with article", "compliance with provisions", "compliance with regulations", "compliance with national", etc.

The same procedure was conducted in order to detect 4-word linguistic forms used for linking with other regulations, which are also similar to the expression "in compliance with". We detected linguistic forms such as "in compliance with law" (which is simultaneously the most frequently occurring 4-word linguistic form), "in compliance with article", "in compliance with national", "in compliance with special", etc. We then repeated the same procedure to detect 5-word linguistic forms used for linking with other laws.

Subsequently, we conducted a procedure of detecting and extracting linguistic forms consisting of a larger number of words (more than 5 words), which are similar to the previously detected linguistic forms. This was also performed through a manual

application of SQL queries in the observed set of regulations, as shown in the following example:

```
SELECT * FROM tblLaws
WHERE ((tblLaws.ArticleText)
Like "*in compliance with*");
```

This way, we were able to detect linguistic forms such as “in compliance with the Constitution”, which is found in 82 articles of law, “in compliance with the Law on Tax Procedure and Tax Administration”, found in 71 articles of law, as well as numerous other linguistic forms used for linking. The overall number of linguistic forms used for linking which were thus detected and extracted amounted to 2,069. The number of links between regulations identified on the basis of these linguistic forms amounted to 38,074. Below, we will show only examples of such linguistic forms, since the detailed list of all detected forms is too large to be included in this document.

4.1. Linguistic forms used for link analysis in legal documents

In their research, Waltl, et al. [31] present the following different reference types in legal documents: Full-explicit reference (FR), Semi-explicit Reference (SR), Implicit Reference (IR) and Tacit Reference (TR). A more detailed analysis of linguistic forms in the observed documents reveals those four types of linking forms (as shown in Figure 1):

1. Linguistic forms used as reference for certain statements of the same or different regulations, which are applied to avoid repetition of certain statements. A statement referring to a separate regulation states the name of that regulation and the issue of the official gazette announcing the publication of the regulation. If it refers to a specific provision within a regulation, the article containing such a provision is also cited [24]. Such linguistic forms are most suitable for link analysis and they belong to the Full-explicit reference (FR) category.
2. Linguistic forms used to refer to a certain law, without specifying any particular article of a law. Such linguistic forms are also suitable for link analysis. This group of linguistic forms belongs to the Semi-explicit Reference (SR) category.
3. Linguistic forms used to refer to a separate regulation by stating a generalized name for the specific type of regulations defining the observed area, without mentioning the exact name of a law or regulation. This group of linguistic forms belongs to the Implicit Reference (IR) category. Such forms are mainly used when it is necessary to follow a chronological order of passing a law in a certain area and respect the hierarchy of acts and application techniques. Some of the linguistic forms contained in this group are:
    - “in compliance with the law regulating %” - such linguistic forms can provide data useful for link analysis, but are not sufficiently accurate. They can be used for a “manual” reference to certain laws.
    - “which regulate the area %” – such forms are also not sufficiently accurate. They can be used to refer to a group of laws belonging to a specific area.

4. Tacit Reference (TR) linguistic forms are those used to generally refer to application of laws (plural), without mentioning a specific law. Such forms do not provide data applicable in link analysis. They include the following examples:
    - "in compliance with law", "in compliance with law and other regulations", with no reference to a specific law, while bearing the general meaning of compliance "with laws".
    - "in compliance with specific regulations"
    - "in compliance with existing regulations"
    - "in compliance with expired regulations valid until", "in compliance with previously valid regulations"
    - "in compliance with valid regulations", "in compliance with existing legal regulations"
    - "in compliance with relating regulations"
    - "in compliance with general regulations"

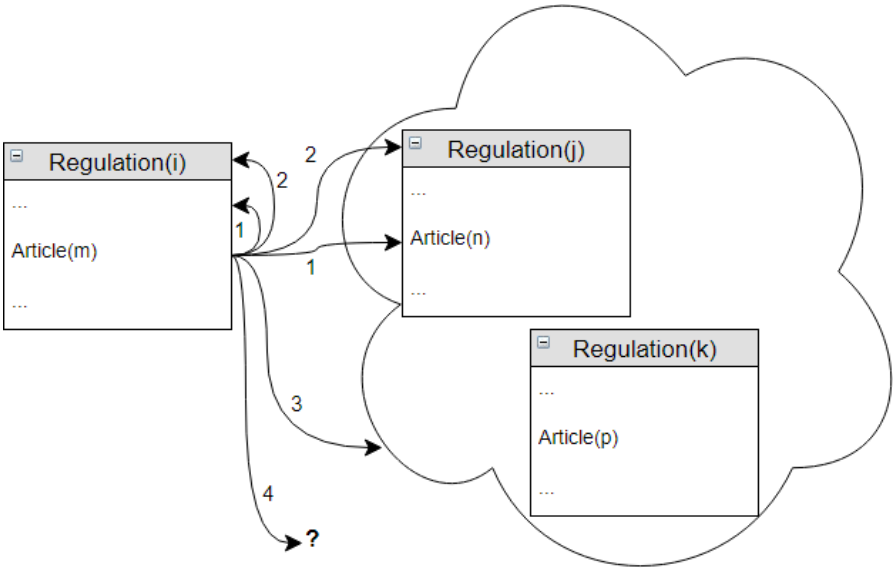

Fig. 1. Types of linking linguistic forms

Most frequently used linking linguistic forms are those used to refer to or link with the same or other paragraphs, articles or laws. They provide the opportunity to perform a detailed referencing and link analysis between laws and parts of laws. Linking linguistic forms can further be divided into the following subgroups:

- **linguistic forms used for auto-referencing or auto-linking**. In the Serbian language, this group recognizes the following linguistic forms:
    - "this article" – most frequently used expression, mentioned in 26.6% of all articles of various laws. This expression is used for auto-referencing or quoting of the same article of the law.

  - "this law" – frequently used expression, mentioned in 22.45% of articles of various laws. This expression is used for auto-referencing or quoting of the same article of the law.
  - "from paragraph % of this article" - frequently used expression, mentioned in 21.2% of articles of various laws. This expression is used to refer to a paragraph of the same article of the law, where "%" stands for the marking or ordinal number of a certain paragraph.
  - "in compliance with this" – frequently used expression, mentioned in 4.68% of articles of various laws. It is used for general auto-referencing or quoting of the text of the same law, article, paragraph etc. For instance, the expression "in compliance with this law" is mentioned in 3.79% of articles of various laws and is used to refer to the same law.
  - "in compliance with article % of this law" (1.29%), where either one or more articles of the law can be mentioned in the text.
  - "based on the provisions stated in article % of this law" (less than 1%)
  - "in compliance with the provisions of this law" (less than 1%), "in compliance with the provisions of this article" (less than 1%)
- **linguistic forms used to refer or link to other national laws, articles or paragraphs**. In the Serbian language, this group recognizes the following linguistic forms:
  - "in compliance with" – frequently used expression, mentioned in 23.66% of articles of various laws. This expression is used for referencing or quoting of a paragraph, article or law.
  - "from paragraph" – frequently used expression, mentioned in 21.67% of articles of various laws. This expression is used for referencing or quoting of a specific paragraph within an article of a law.
  - "in compliance with the law" - frequently used expression, mentioned in 6.10% of articles of various laws. This expression is used for referencing or quoting of a specific law.
  - "in compliance with the Constitution" (less than 1%)
- **linguistic forms used to refer or link to international laws or international institutions** (Note: this research has not specified the exact international laws bearing links and references with domestic laws, but rather identified the existence of these connections):
  - "Law on Validation %", "Law on Ratification %" (less than 1%)
  - "harmonization % with the respective regulations of the European Union", "between the Republic of Serbia and the European Union", "EU laws", "the laws of EU", "in the Official Journal of the European Union", "competent authorities of the European Union", "The process of stabilization and association with the European Union", "in compliance with % regulations of the European Union" (less than 1%)
  - "in compliance with international obligations", "in compliance with internationally ratified agreements", "in compliance with the international agreement concluded between", "in compliance with generally recognized regulations of international law" (less than 1%)
  - "in compliance with the Resolution of the Security Council of the United Nations", "in compliance with the Resolution of the Security Council of

UN", "in compliance with the Resolution #1244 of the Security Council of the United Nations" (less than 1%).

- **linguistic forms used to refer or link to lower-level legal acts** (such expressions refer only to the existence of lower-level legal acts. This paper has not analyzed respective links and references):
  - "in compliance with the agreement", "in compliance with that agreement" - expressions, mentioned in about 1% of articles of various laws.
  - "in compliance with the statute" (less than 1%)
  - "in compliance with special acts", "in compliance with the act" (less than 1%)
  - "in compliance with the instructions" (less than 1%).

When exploring linguistic forms and detecting those suitable to serve as links for the purposes of our research, we have structured the links so that they are consisted of two sections:

- law ID and
- law article ID

According to this structure, a link relating Article (m) of Regulation (i) to Article (n) of Regulation (j) can be presented in the following manner:

$$\text{Link(Source)=Target}$$

$$\text{Link(Regulation}(i)\text{,Article}(m)\text{)=[Regulation}(j)\text{,Article}(n)\text{]}$$

In cases of linguistic forms used to make reference to a law, without stating a specific article of the law, we created links consisting of law IDs, while replacing law article ID with ZERO:

$$\text{Link(Regulation}(i)\text{,Article}(m)\text{)=[Regulation}(j)\text{,0]}$$

This way, we acquire a model consisting of a set of entities and a set of relations between them, while most operations conducted over such sets are modeled as operations in graphs [5].

The text of Regulation (*i*) can have multiple references to Regulation (*j*), but our research has not considered how many times a Regulation (*j*) is mentioned in Regulation (*i*). In our further research, we have only used retrieved information confirming the existence of links between certain regulations, while disregarding the number of times the text of a Regulation (*i*) mentions a Regulation (*j*). Therefore, in our research, connections and links between two regulations are *unweighted* and the graph so given is *unweighted* [6].

Evaluating legal documents link extraction based on detected linguistic forms. In our research, we firstly conducted the extraction of all linguistic forms from observed sets of regulations and then focused on those which can be utilized to link with or refer to the same or different laws. This paper, however, provides result assessment of only those linguistic forms which can be utilized for linking.

According to Miner, et al. [17], the efficacy of an entity extractor on a corpus is determined by scoring the output of the system against known labels for the same type of corpus. Therefore, we extracted a test set of regulations, which was established to contain 3,111 links to the same regulation or other regulations. This set was then used to test the above mentioned efficacy. Entity extraction systems are scored using precision

and recall. The precision of an extractor is the percentage of predicted named entities that are correct:

$$\text{Precision}(p)=\#\text{correct}/\#\text{found} \quad (1)$$

Recall is the percentage of occurrences of a given named entity found by the system compared to total occurrences of the entity found in all of the data:

$$\text{Recall}(r)=\#\text{correct}/\#\text{in true data} \quad (2)$$

Usually, precision and recall are combined in a single measure called the "F-measure". That is a weighted average of precision and recall. Most often, precision and recall are weighted equally, causing the following equation [17]:

$$f1 = (2*p*r) / (r+p) \quad (3)$$

Bearing in mind that this research firstly conducted an extraction of all linguistic forms and then separated and examined only those used for linking, then, according to equation (1), the value of the precision of an extractor is:

$$\text{Precision}(p)=1$$

In order to show more clearly how the maximum value of this parameter was acquired, let us recall that the initial stage of our research provided a collection of all linguistic forms, which implies that the linking forms were certainly contained in the texts of regulations. From this collection, we then manually extracted those forms which we established were used for linking, in the manner described in Section 4. Therefore, within the detected linking linguistic forms there are no forms which do not occur in the texts of regulations and none which are not used for linking. Consequently, the parameter *p* has a maximum value.

On the other hand, based on our data and according to equation (2), the acquired value of Recall is:

$$\text{Recall}(r)=0.9955$$

Within the observed set of regulations, for which it had been established to contain 3,111 links, we successfully detected 3, 097 links. These links were detected through the linking linguistic forms identified in the manner previously described. The test set of regulations also contained 14 undetected links between regulations, which is an error of 0.45%. Table 3 presents linguistic forms which were not detected in the observed set of regulations, the number of their occurrences and their respective links to other laws.

Table 3. Undetected linguistic forms use for link with other laws, the number of their occurrences in the observed set and respective laws

Linguistic Form	Number of occurences	Link to the law
"Bankruptcy Law"	1	The Law on Bankruptcy
"of the civil litigation procedure"	1	The Law on Litigation

Linguistic Form	Number of occurences	Link to the law
"Interim Agreement on trade and trade-related matters"	1	The Law on the Ratification of the Interim Agreement on trade and trade-related matters between the European Community, of the one part, and the Republic of Serbia, of the other part
"if the patient is deaf-mute"	2	The Law on the Use of Sign Language
"the right to confidentiality of all personal data"	1	The Law on Personal Data Protection
"the right of insight into medical documentation"	1	The Law on Medical Documentation and Records in the Field of Healthcare
"discrimination on the grounds of mental disability"	1	The Law on the Prohibition of Discrimination
"considered personal data"	2	The Law on Personal Data Protection
"of interest to public health and security"	1	The Law on Public Health
"Information from medical records and documentation"	2	The Law on Medical Documentation and Records in the Field of Healthcare
"Medical records and documentation"	1	The Law on Medical Documentation and Records in the Field of Healthcare
	Total: 14	

Based on these values, and in accordance with equation (3), the "F-measure" value can also be calculated as:

$$f1 = (2*1*0.9955) / (0.9955+1) = 0.9977$$

Based on the retrieved values, it is determined that all the detected entities are at the same time correct, with the ratio of the number of identified entities and the overall number of entities not exceeding 100%. In the group of linguistic forms which have a high rate of occurrence in the observed set of regulations we manually recognized and separated linguistic forms which are used for referencing and linking. It is important to point out that some linking linguistic forms which have a low rate of occurrence have certainly been left out in this procedure. As a result, the obtained Recall value is lower than 100%. This value can be increased through a more precise search and extraction of linking linguistic forms with a low frequency of occurrence.

4.2. Analysis of Paths Between Related Laws

According to the theory of graphs [4], each graph consists of nodes, or vertices, where the nodes may or may not be mutually connected. The links connecting two nodes are

lines, or edges, and can be directed or undirected. In our experiment, it is obvious that links between regulations are directed, which makes the graph they form directed. In directed graphs, in-degrees and out-degrees must be determined, where the in-degree of a node is the number of incoming edges ending at the observed node and its out-degree is the number of outgoing edges starting from the mentioned node [6].

Our further analysis focuses on paths within the observed set of regulations. A path from *i* to *j* in graph *G* is a sequence of distinct vertices starting with *i* and ending with *j* such that consecutive vertices are adjacent [1]. In a directed graph, a directed path (sometimes called dipath) is again a sequence of edges which connect a sequence of vertices, but with the added restriction that the edges all be directed in the same direction [23]. Our target was to detect all simple paths within a set of all observed regulations. A "simple path" is the one which does not contain recurrent nodes, that is, all the nodes are different. In other words, a simple path has each of the nodes appearing only once [6].

In order to achieve this, our model uses all the identified links between regulations, including links to articles and paragraphs of law, being observed as references to a particular law while disregarding which article or paragraph of the mentioned law the references are pointed at. The following algorithm has been used to detect all simple paths:

```
for all node_source in Graph:
  for all node_target in Graph:
    find all simple paths(Graph) from node_source to
node_target
    insert into table all simple path(Graph)
```

Having applied the described procedure, we managed to identify 210,230 different paths between graph nodes, i.e. different regulations. The majority of paths begin with the Law on Hibernation and Cancellation of Debts on Compulsory Health Insurance Taxes, while the majority of paths lead towards international laws, with the Law on General Administrative Procedure in second place.

Generally speaking, in cases of any two nodes *i* and *j*, within a directed graph *G*, we can define the distance between those nodes as the shortest path from node *i* to node *j* [4]. However, in the model presented in our research, we were particularly interested in detecting the longest path, or diameter of an observed graph. A diameter of a graph *G* is the longest distance between two nodes, which provides information on the two most distant nodes within a graph. The longest paths we have been able to detect in our model contained 45 nodes, or more precisely, 45 different, mutually related laws.

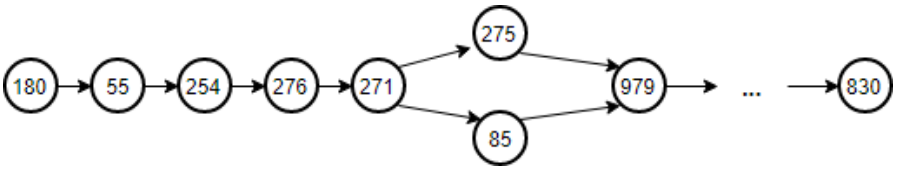

Fig. 2. Graphic representation of the two longest paths detected in the observed set of regulations

The observed set of regulations contains two different longest paths (Figure 2), and both share the starting and finishing laws: the starting law is the Law on Hibernation

and Cancellation of Debts on Compulsory Health Insurance Taxes (in our model, the ID of this law is 180), while the finishing law is the Law on Transformation of Agricultural Public Property into Other Types of Property (ID 830). These two paths differ only in as much as in the first case the path leads from the Law on Tax Procedure and Tax Administration (271), over the Value-Added Tax Law (275), towards The Law on Tourism (979), while in the other case, the path starts at the Law on Tax Procedure and Tax Administration (271), over the Law on Local Self-Government Financing (85), towards the Law on Tourism (979).

4.3. Detection of "High-prestige" Legal Documents, based on their Correlation

Many different measures of prestige and centrality of a node have been defined in Graph Theory, and some authors call these measures "importance" "standing" "prominence" or "popularity" [6]. When the graph is directed, we speak about prestige and importance (they are used interchangeably).

Generally speaking, a degree of confidence of node *i* in node *j* is the number of references leading from node *i* to node *j* [6]. Based on the example of information on links between laws in the observed set of laws, we have defined the degree of confidence in law *j* as the overall number of laws linked with the observed law *j*. Therefore, the In-degree of confidence is the number of incoming links from other laws, while the Out-degree of confidence is the number of links outgoing from that law towards other laws.

Most laws have a low In-degree and Out-degree of confidence, that is, there is a small number of incoming and outgoing links in these laws. On the other hand, it is noticeable that certain laws have a more significant In-degree and Out-degree of confidence. The highest In-degree of confidence is noticed in the Constitution of the Republic of Serbia, which, according to the data available to us, has 47 incoming links from other laws. The highest Out-degree of confidence is found in the Law on Tax Procedure and Tax Administration, which shows 25 outgoing links.

Numerous algorithms have been developed to measure levels of node prestige in a directed graph, most commonly by gauging the number of citations or confirmations (using the number of incoming links) towards a given node. According to Fouss, et al. [6], the most popular algorithms to measure the level of node prestige in a directed graph, which are predominantly used in social sciences, are the algorithms from the Classic Node Prestige Measures group, such as: Node In-degree, Prestige by Proximity, A Spectral Measure of Prestige and Prestige Based on Indirect Links. According to the same source, other algorithms, which were developed mainly for the purposes of bibliometrics and use with Internet search engines, include: Citation Influence, Rating Methods Based on Least Squares, PageRank Algorithm, HITS: Hubs and Authorities, Probabilistic HITS and A Simple Bag-of-Paths Prestige Measure.

In order to detect high-prestige legal documents, based on their mutual links, our research and its findings suggested the use of algorithms primarily applied in social sciences.

Node In-degree is the most basic measure of node prestige within a graph and corresponds to the number of lines incoming towards a node. The highest prestige, according to the application of this algorithm on our set of data, is acquired by the Constitution of the Republic of Serbia (1.00), followed by the Law on General

Administrative Procedure (0.66), the Law on Economic Associations (0.64), the Law on Amendments to the Law on Fines for Business Violations and Infringements (0.55), the Law on Accountancy (0.45), the Law on Personal Data Protection (0.43), The Law on State Government (0.40), the Law on Healthcare (0.40), the Law on Planning and Construction (0.40), the Law on Budgetary System (0.36), the Law on Obligations (0.34), and so on. The values given in parentheses are relative values acquired according to the described algorithm, where the maximum value can equal 1, while all other values are determined in correspondence to the maximum value.

Prestige by Proximity is the algorithm measuring the degree at which a node within a graph is directly or indirectly (via an intermediary) cited by other network nodes, where "the proximity prestige" quantifies the importance of the region of influence of a node in the directed graph [6]. To compute the proximity prestige of a node *j*, the first step is to calculate the cost of the shortest path, between all pairs of nodes *i* and *j*. According to this algorithm, a node *j* will attain a higher prestige index in case a large number of network nodes directly or indirectly cite node *j*, where the average distance of those nodes to node *j* is short. The highest prestige, according to the application of this algorithm on our set of data, is acquired by the Law on the President of the Republic of Serbia (1.00), the Law on Local Self-government (1.00), the Law on the Government of the Republic of Serbia (1.00), the Constitution of the Republic of Serbia (0.70), the Law on Inheritance (0.64), the Law on Foreign Trade (0.64), the Law on the Market of Capital (0.64), the Law on Litigation (0.64), and so on. The values given in parentheses are relative values acquired according to the described algorithm.

A Spectral Measure of Prestige is a measure of prestige, adapted here to directed graphs, stating that the prestige of node *j* is proportional to the sum of the connection weights from all the nodes citing *j*, multiplied by the prestige of these nodes citing *j* [6]. According to this algorithm, a node *j* will attain a higher prestige index in case a node *j* has been cited by a significant number of other nodes and if the nodes citing node *j* also bear high prestige. The idea behind this algorithm is quite similar to the concept of PageRank algorithm, which was later developed to rank web pages. The highest prestige, according to the application of this algorithm on our set of data, is acquired by the Constitution of the Republic of Serbia (1.00), followed by the Law on General Administrative Procedure (0.66), the Law on Economic Associations (0.64), the Law on Accountancy (0.47), the Law on Personal Data Protection (0.43), the Law on Planning and Construction (0.40), the Law on Healthcare (0.40), and so on.The values given in parentheses are relative values acquired according to the described algorithm.

Prestige Based on Indirect Links is an algorithm (or a set of algorithms) which measures the degree of graph node prestige taking into account direct and indirect links [6]. According to this algorithm, the prestige of a graph node is most influenced by direct links, followed by the lesser influence of second-step indirect links, third-step indirect links and so on. This algorithm is strongly affected by the factor of reduction α (where $0<\alpha<1$), which measures the reduction of indirect links from k-step, by multiplying the prestige of these links with α^k. When the factor of reduction is large, the influence of long paths is only slightly diminished, and the result tends to be related with the result attained through "A Spectral Measure of Prestige" algorithm. Vice versa, when the factor of reduction α borders zero, the influence of long paths on a node abruptly declines and the acquired result tends to be related with the result attained through "Node In-degree" algorithm. As the application of this algorithm while using border values for the factor of reduction α tends to yield results similar to those from the

previously described algorithms, this research has not used the mentioned algorithm. The influence of indirect links α importance reduction factors on measuring the importance of legal documents based on their mutual correlation can be the subject of future research.

In addition to using algorithms to detect high-prestige nodes in directed graphs, which is a method predominantly used in social sciences, our research has also used **Citation Influence** algorithm [6], which is mainly utilized in bibliometrics in order to establish the measures of prestige in magazines. The basic idea behind this algorithm is to measure the balance of incoming and outgoing links and citations within every node in a graph. According to this algorithm, an indicator of the efficiency of some node *j* is the ratio of volume of input citations to *j* and volume of output citations from *j* [6].

The highest prestige, according to the application of this algorithm on our set of data, is acquired by the Law on General Administrative Procedure (1.00), followed by the Law on State Government (0.61), the Law on Accountancy (0,34), the Law on Administrative Proceedings (0.29), the Law on High Education (0.26), the Law on Occupational Safety and Healthcare (0.26), the Constitution of the Republic of Serbia (0.25), and so on. The values given in parentheses are relative values acquired according to the described algorithm.

According to the gathered data, we have concluded that different algorithms applied in order to measure the prestige of a legal document based on mutual correlation between laws within an observed collection of regulations yield different results. However, bearing in mind that the Constitution represents the very "summit" of one country's legal system, standing as the foundation for all other laws, we find that "Node In-degree" and "A Spectral Measure of Prestige" confirm this conclusion, while other algorithms provide different results.

Legal science determines a clear hierarchy among regulations, upon which all laws bear equal importance, and argues against allocating certain laws with higher or lower importance. Therefore, our evaluation of the effects of the mentioned algorithms has been conducted only in relation to the Constitution. However, through analysis of links between laws and the application of the mentioned algorithms it is possible to determine possible effects arising from amendment to certain laws, when such amendments are necessary.

5. Discussion and Future Work

The purpose of this research was not to interpret laws, nor have we attempted to delve into legal science. This research has incorporated the usage of certain techniques and algorithms from data science in order to extract information from a collection of legal documents.

Based on the example of a set of laws applicable in a single country, we have explored and identified all the linguistic forms used when compiling laws. We have extracted all the said linguistic forms and established the frequency of occurrence of these forms appearing in the observed texts, in an effort to further analyze acquired information. In this manner, we obtained accurate data on the extent at which certain linguistic forms are used when compiling texts in the observed set of laws. It is immediately noticeable that the most frequently occurring linguistic forms are used for

referencing and linking. Assessment of results attained through the extraction of linking linguistic forms has proved that "The Precision of an Extractor" parameter yielded excellent results. In contrast, the "Recall" parameter has not provided the maximum results, which means that not all linking linguistic forms have been identified. This value can be increased through a more precise search and extraction of linking linguistic forms with a low frequency of occurrence.

Based on the attained information, we have proceeded to analyze links and paths within the observed set of regulations. Apart from establishing which laws have the highest numbers of incoming and outgoing links, we have also identified the longest paths which, in the case of our study, consist of 45 different mutually related laws.

Such information is particularly significant when it is necessary to amend certain laws. Amending a law which is contained in a graph node influences not only the area regulated by that law, but can also affect laws and areas linked with it and belonging to the same path. For instance, the example of the law situated at the end of the longest path suggests that its amendments can affect 44 other laws and areas regulated by those laws.

The existence of a system of links between legal documents is particularly important when it is necessary to amend such documents. Such amendments, or more precisely alterations and amendments, are conducted when a regulation needs to be harmonized with alterations of the legal system, alterations in policies affecting a certain area, or actual practice. Principally speaking, altering and amending certain regulations does not automatically alter and amend other regulations, but enables establishing the cessation of validity of certain stipulations contained in other regulations [24]. Therefore, adequate insight into the effects of alterations of a certain regulation to other regulations is of utmost importance. It is safe to assume that every regulation refers or links to other regulations. Consequently, alterations of a regulation affect all related regulations. Establishing a system of links between legal documents, laws and other regulations ensures an efficient management of alterations to these documents, taking into account their interdependence.

The legal science determines a clear hierarchy among regulations, upon which all laws bear equal importance. But some necessary amendments can possibly cause effects to certain laws. By using an analysis of links between laws and the application of algorithms designed to identify high prestige graph nodes, these effects can be determined.

The application of Graph Theory allows for the detection of high-prestige nodes within a graph. We applied several algorithms on the material collected in our research in order to identify legal documents of higher prestige on the basis of their mutual correlation. The findings confirmed that different algorithms yielded different results. Our research has concluded that, from the standpoint of the role which the Constitution plays in a legal system, "Node In-degree" and "A Spectral Measure of Prestige" algorithms provide expected results, unlike other algorithms.

As far as the practical application of automatic management of complex information in legal documents is concerned, one of the approaches suggests standardization of technical preparation or codification of all new regulations. For instance, certain existing meta-schemas for semantic representation of legal documents, such as "Akoma Ntoso" [2], can be used for such purposes. Such practice would facilitate and simplify management of alterations within the hierarchy of regulations, bearing in mind their mutual interdependence. As far as the existing legal documents are concerned, the

application of techniques belonging to the text mining or information extraction domain would ensure technical preparation of such documents in order to enable a more advanced form of their automatic exchange and connection to satisfy the needs of the wider community.

Linguistic forms mining in legal documents shows that the most frequently used linguistic forms are used to refer or link to the same law or certain other laws. However, there are also many other linguistic forms. Future research can proceed in two directions.

The first direction of future research would involve further link analysis of identified linguistic forms. This study has observed only links between certain laws, without analyzing links towards law articles and paragraphs. Future research should certainly include these issues as well. Furthermore, our research has examined only texts of laws. Future research could and should also consider other regulations and lower-level legal acts. Legal texts commonly contain guidelines defining [24]: the relation between the regulation which ceases to be valid and the new regulation in terms of the effects they have on matters, situations and correlations created during the period of validity of the previous regulation; acting in pending cases; deadlines and authorization issues regarding the adoption of by-law regulations; the relation between a new regulation and by-law regulations adopted in compliance with the previously valid regulation and the necessity to amend previous and adopt new by-law regulations in compliance with the new regulation; the retroaction incurred in compliance with certain law provisions; information on specific time-related restrictions when applying a regulation (time limiting provisions). This implies that it is possible to conduct further link analyses between law and by-law regulations in order to establish both a hierarchy of regulations and determine relations with previously valid regulations and with new regulations.

The second direction of future research would involve extraction of information from legal documents in the manner described by Feldman & Sanger [5]. The linguistic forms detected through our research also include those which enable the extraction of information from legal documents, as well as linguistic forms related to time guidelines, events, other features (metadata) of legal documents and so on.

The extraction and analysis of linguistic forms in legal documents also enables identification of events, which are, according to Feldman & Sanger [5], some of the basic elements attained through text extraction. These events include the date of a law coming into action, the date of publication, the date of the commencement of the application of a law, the date of the cessation of validity of a law and so on. Garofalakis, et al. [8] deal with the issue of different law versions in their work. In practice, a regulation officially comes into action upon expiration of a certain period after its publication. Sometimes, the date of the commencement of the validity and the date of the commencement of the application of a law, or some of its provisions, may be separated by a period of time. Also, the date of a law coming into action and the date of the commencement of its application may be separated by a period of time [24]. These are some additional linguistic forms detected in the course of our research which refer to time guidelines and can be made subject of future work.

Additional identified linguistic forms which enable “extraction of information” from legal documents, extraction and analysis of relevant facts and similar features include: linguistic forms defining the rights and obligations of legal entities, linguistic forms defining authorization issues (provisions on by-law regulations which are to be passed in order to ensure enforcement of laws), linguistic forms defining penal provisions

(ordering and prohibiting norms), linguistic forms proscribing time periods of a legal status taking place etc.

For the purposes of automatic information management in legal documents, the extraction and analysis of linguistic forms described in this paper can be used to identify features and metadata in legal text documents. The issue of recognizing the structure and metadata of legal documents from plain text and modeling them according to the meta-schema used for the semantic representation of legal resources was the subject explored by Koniaris, et al. [13] in their study. The method they described exploits common formats of legal documents to identify blocks of structural and semantic information and model them according to a popular legal meta-schema. Any document, including legal documents, can possess a large number of metadata. The metadata of legal documents is commonly found in the final section of the text of the regulation and can involve transitional and concluding provisions. The metadata related to legal documents include: the date of passing a regulation, the number at which a regulation was published or recorded, the authority which passed the regulation, the authorized person who signed the regulation and so forth. The linguistic form most commonly used to enable identification of metadata on where a law is published is “this law has been published in %”, where the marking “%” stands for the name of the public newsletter where the law was published.

6. Conclusion

This research has shown the application of certain data science techniques and algorithms in order to extract information from a collection of legal documents, using statistical approach in extracting linguistic forms. Based on the example of a set of laws applicable in a single country, we have explored and identified all the linguistic forms used when compiling laws. In this manner, we have obtained accurate data on the extent at which certain linguistic forms are used when compiling texts in the observed set of laws. The methodology to linguistic forms extraction, herein described, can be utilized not only in the mentioned case, but rather on any collection of texts applicable in other languages.

In the manner described for the observed set of regulations, we have shown that it is possible to detect and create references for texts of laws and other regulations. Thus obtained information can be used to explore legal documents, analyze references and links in legal documents, analyze paths between related legal documents and measure the level of importance of legal documents on the grounds of their mutual interdependence. The information gathered in this manner can prove useful when evaluating the effects of possible law amendments on all related laws and areas of interest.

The number of different ways in which exploration of linguistic forms in legal documents can be used certainly exceeds the few examples here stated, while the information attained in such manner forms an excellent basis and acts as input data for numerous further analyses.

References

1. Avis, D., Hertz, A. & Marcotte, O., 2005. In: *Graph Theory and Combinatorial Optimization*. s.l.: Springer US, pp. 3-6.
2. Barabucci, G. et al., 2009. Multi-layer Markup and Ontological Structures in Akoma Ntoso. In: *AI Approaches to the Complexity of Legal Systems. Complex Systems, the Semantic Web, Ontologies, Argumentation, and Dialogue*. s.l.: Springer, Berlin, Heidelberg, pp. 133-149.
3. Bommarito II, M. J. & Katz, D. M., 2009. *Properties of the United States Code Citation Network*, s.l.: SSRN´s eLibrary.
4. Bondy, A. & Murty, U., 1982., Graphs and Subgraphs; Directed Graphs. In: *Graph Theory with Applications*. s.l.: NORTH-HOLLAND, pp. 1-20, 171-178.
5. Feldman, R. & Sanger, J., 2006. Introduction to Information Extraction; Link Analysis; Information Extraction in the DIAL Environment. In: *The Text Mining Handbook*. s.l.: Cambridge University Press, pp. 94-96, 244-274, 317-318.
6. Fouss, F., Saerens, M. & Shimbo, M., 2016. Basic Graph Concepts; Identifying Prestigious Nodes. In: *Algorithms and Models for Network Data and Link Analysis*. s.l.: Cambridge University Press, pp. 7-11, 201-207.
7. Furlan, B., Batanović, V. & Nikolić, B., 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3), pp. 710-719.
8. Garofalakis, J., Plessas, K. & Plessas, A., 2016. A Semi-automatic System for the Consolidation of Greek Legislative Texts. In: *Proceedings of the 20th Pan-Hellenic Conference on Informatics*. Patras, Greece: ACM, pp. 1:1-1:6.
9. Gillam, R., 2003. Case Folding. In: *Unicode Demystified: A Practical Programmer's Guide to the Encoding Standard*. s.l.: Addison-Wesley, p. 588.
10. Kajan, E., Pljasković, A. & Crnišanin, A., 2012. *Normalizacija tekstualnih dokumenata na sprskom jeziku u cilju efikasnijeg pretraživanja u sistemima e-uprave*. Zlatibor, ETRAN.
11. Karanikolas, N. N. & Skourlas, C., 2006. Text Classification: Forming Candidate Key-Phrases from Existing Shorter Ones. *FACTA UNIVERSITATIS*, 19(3), pp. 439-451.
12. Karanikolas, N. N. & Skourlas, C., 2010. A parametric methodology for text classification. *Journal of Information Science*, 36(4), pp. 421-442.
13. Koniaris, M., Papastefanatos, G. & Vassiliou, Y., 2016. *Towards Automatic Structuring and Semantic Indexing of Legal Documents*. Patras, Greece, PCI'2016: Proceedings of the 20th Pan-Hellenic Conference on Informatics.
14. Liu, J., Shang, J. & Han, J., 2017. Quality Phrase Mining with User Guidance. In: *Phrase Mining from Massive Text and Its Applications*. s.l.: Morgan & Claypool, pp. 5-34.
15. Lodder, A. R. & Oskamp, A., 2006. Drafting and traditional retrieval. In: *Information Technology and Lawyers: Advanced Technology in the Legal Domain, from Challenges to Daily Routine*. s.l.: Springer Science & Business Media, pp. 144-147.
16. Milošević, N., 2012. Stemmer for Serbian language. *CoRR abs/1209.4471*.
17. Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., Delen, D., 2012. The Seven Practice Areas of Text Analytics; Conceptual Foundations of Text Mining and Preprocessing Steps; Entity Extraction. In: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. s.l.: Academic Press, pp. 31-32, 45-50, 924-928.
18. Monroy, A. L., Calvo, H., Gelbukh, A. & Pacheco, G. G., 2013. Link Analysis for Representing and Retrieving Legal Information. In: A. Gelbukh, ed. *Computational Linguistics and Intelligent Text Processing*. s.l.: Springer Berlin Heidelberg, pp. 380-393.
19. Neale, T., 2013. Citation Analysis of Canadian Case Law. *Journal of Open Access to Law*, 1(1).
20. Opijnen, M. v., Verwer, N. & Meijer, J., 2015. *Beyond the Experiment: The Extendable Legal Link Extractor*. s.l., International Conference on Artificial Intelligence and Law (ICAIL).

21. Oxford University Press, 2017. *English Dictionary, Thesaurus, & grammar help | Oxford Dictionaries*. [Online] Available at: https://en.oxforddictionaries.com/ [Accessed 22 05 2017].
22. Prince, V. & Labadié, A., 2007. Text segmentation based on document understanding for information retrieval. In: Z. Kedad, et al. eds. *Natural Language Processing and Information Systems*. s.l.: Springer Berlin Heidelberg, pp. 295-304.
23. Robertson, N. & Seymour, P. D., 1993. *Graph Structure Theory*. s.l., American Mathematical Society.
24. RS, Z. o. N. s., 2010. *Jedinstvena metodološka pravila za izradu propisa*. s.l.: Službeni glasnik RS, 21/2010.
25. Sakhaee, N., Wilson, M. C. & Zakeri, G., 2016. New Zealand Legislation Network. In: *Legal Knowledge and Information Systems*. s.l.: IOS Press, pp. 199-202.
26. Saravanan, M., Ravindran, B. & Raman, S., 2008. *Automatic Identification of Rhetorical Roles using Conditional Random Fields for Legal Document Summarization*. s.l., International Joint Conference on Natural Language Processing.
27. Stranieri, A. & Zeleznikow, J., 2005. Legal issues in the data transformation phase; Information retrieval and text mining. In: *Knowledge Discovery from Legal Databases*. s.l.: Springer, pp. 59-84, 147-170.
28. Vasiljević, N., 2015. *Automatic Processing of Legal Text in the Serbian Language: doctoral dissertation*. [Online] Available at: https://phaidrabg.bg.ac.rs/detail_object/o:10687 [Accessed 05 10 2017].
29. Vitas, D. et al., 2012. Particularities of the serbian language; Other application areas. In: *The serbian language in the digital age*. s.l.: Springer, Berlin, Heidelberg, pp. 49-54, 66-67.
30. Wagh, R. S., 2014. Exploratory Analysis of Legal Documents using Unsupervised Text Mining Techniques. *International Journal of Engineering Research and Technology*, 3(2).
31. Waltl, B., Landthaler, J. & Matthes, F., 2016. *Differentiation and Empirical Analysis of Reference Types in Legal Documents*. Sofia Antipolis, France, Jurix: International Conference on Legal Knowledge and Information Systems.
32. Zimmermann, F., 2010. *Dublin Core and legal informatics, VoxPopuLII, jurMeta - New Metadata Initiative for Legal Documents*. [Online] Available at: https://blog.law.cornell.edu/voxpop/category/dublin-core-and-legal-informatics/ [Accessed 25 01 2017].

Đorđe Petrović is a PhD candidate in Computer science and Informatics at the Faculty of Electronic Engineering, University of Niš. He has more than 18 years of experience working mainly with web technologies. His current research interests include area of text mining and machine learning. He is teacher at Department of Business Informatics at Valjevo Business School of Applied Studies.

Milena Stanković received the B.Sc. degree in electronic engineering from the faculty of Electronic Engineering University of Nis, Serbia, in 1976, and M.Sc. and Ph.D. degrees in Computer science and Informatics from the Faculty of Electronic Engineering, University of Nis, in 1982 and 1988, respectively. Currently, she is a Head of the Department of Computing, Faculty of Electronic Engineering and Head of CIITLab (Computational Inteligence and Information Technologies Laboratory). Her research interests include switching theory, multiple-valued logic, spectral techniques and data mining.

Received: July 1, 2017; Accepted: March 20, 2018.