Automatic Identification of Legal Terms in Czech Law Texts

Karel Pala, Pavel Rychlý, Pavel Šmerk

Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{pala,pary,xsmerk}@fi.muni.cz

Abstract

Law texts including constitution, acts, public notices and court judgements form a huge database of texts. As many texts from small domains, the used sublanguage is partially restricted and also different from general language (Czech). As a starting collection of data the legal database Lexis containing approx. 50,000 Czech law documents has been chosen. Our attention is concentrated mostly on noun groups which are the main candidates for law terms. We were able to recognize 3992 such different noun groups in the selected text samples. The paper also presents results of the morphological analysis, lemmatization, tagging, disambiguation, and the basic syntactic analysis of Czech law texts as these tasks are crucial for any further sophisticated nanural language processing. The verbs in legal texts have been explored preliminarily as well. In this respect we are trying to explore how the linguistic analysis can help in identification of the semantic nature of law terms.

1. Introduction

In the paper we describe the first results of the new project whose final goal is to build an electronic dictionary of Czech law terms. We have started with a legal database Lexis developed at the Institute of Law, Czech Academy of Sciences, which presently includes approx. 50,000 Czech law documents ranging from the beginning of Czechoslovak Republic in 1918 to present days. It also includes court judgements, main representative law textbooks and law reports. All the texts exist in electronic form.

The first part of the paper presents results of the preparation step for the subsequent term identification – the morphological analysis. For this purpose we have used the tools developed in the Natural Language Processing Centre of the Faculty of Informatics, Masaryk University, particularly, morphological analyser Ajka (Sedláček, 2005) performing lemmatization and tagging and a new tool for grammatical disambiguation named DESAMB (Šmerk, 2007). The tools have been designed for general Czech but it appears that they can be exploited for law sublanguage with some minor modifications, namely adding law terms. The tools are now configured to analyze all Czech law texts contained in the Lexis database, the presented results from the pilot project are described below.

In the second part we report about term identification via syntactic analysis which has used tool DIS/VADIS (Žáčková, 2002), a partial parser for Czech. As a result list of noun groups has been obtained that can be considered as good candidates for law terms. We are also having a look at the verbs existing in law texts because they are relational elements linking together the established law terms. Here the apparatus of valency frames (Horák and Hlaváčková, 2005) comes as an appropriate instrument. It allows us to explore context patterns in which law terms occur and see how they behave in the law text.

The general goal is to find out in what extent linguistic analysis can contribute to semantic analysis of the law text. We are at the beginning of this enterprise.

1.1. Pilot project

As a pilot project we have decided to analyse the current version of the Penal Code of the Czech Republic. It is one of the biggest law documents containing almost 36,000 word forms. The overall characteristic of the document can be found in Table 1.

Number of	
word forms (tokens)	35,893
numbers	2,647
punctuation marks	9,135
tokens total	47,865
different word forms (types)	5,019
different numbers	467
different punctuation marks	12
types total	5,019

Table 1: The overall characteristic of the Penal Code of the Czech Republic

The task is to process the document by the Czech morphological analyser (lemmatizer) Ajka in such a way that for each word form in the source text a morpho-

logical information in the form of morphological tags is obtained. Thus we get information to what parts of speech the word forms belong, and, for instance, for nouns also grammatical categories like gender, number and case. Each word form in the document is associated with its respective lemma as well. In the highly inflectional language like Czech all this information is relevant for the further analysis of law terms. The results of the morphological analysis and lemmatization are transformed into a special format which is described below.

2. Morphological Analysis

We have used several simple scripts to create what is called vertical file from the source text. It is a plain text file without any formatting (word-processing options). Words are written in a column, i.e. each line contains one word, number or punctuation mark. Optional annotation is on the same line and the respective words are divided by the tabulator character. The first step uses only word forms from the source text. The vertical file serves as an input text for many corpus processing tools like CQP (Schulze and Christ, 1996) and Manatee (Rychlý, 2000).

In the next step, we processed the vertical file with the morphological analyser Ajka (Sedláček, 2005). It is a tool exploited for annotating and lemmatizing general Czech texts, however, the processing law texts requires modifications, e.g. enriching the list of stems of Ajka. The programme yields all possible combinations of lemma and morphological tags for each Czech word form.

Table 2 presents an example of the Ajka output, the tag **k1gFnSc1** means: part of speech (**k**) = noun (**1**), gender (**g**) = female (**F**), number (**n**) = singular (**S**) and case (**c**) = first (nominative) (**1**), tags beginning with **k2** are adjectives, **k3** – pronouns, **k5** – verbs and **k7** – prepositions.

As one can see, many word forms are ambiguous: there are more than one possible tag or even lemma for a given word form. In the analysed document, 76 % of word forms are ambiguous, more than 42 % of word forms have more than one possible lemma and average number of tags for an ambiguous word form is 6.75.

We have used part-of-speech tagger Desamb (Šmerk, 2007) to disambiguate such word forms. The output of the Desamb tool contains only the most probable lemma/tag for each word form. Table 3 contains output of Desamb for the input text above.

The annotated version of the document contains 2,560 different lemmas. Frequencies of each part of speech are in Table 4.

Příprava	příprava	k1gFnSc1
k	k	k7c3
trestnému	trestný	k2eAgInSc3d1
činu	čin	k1gInSc3
je	být	k5eAaImIp3nS
trestná	trestný	k2eAgFnSc1d1
podle	podle	k7c2
trestní	trestní	k2eAgFnSc2d1
sazby	sazba	k1gFnSc2
stanovené	stanovený	k2eAgFnSc2d1
na	na	k7c4
trestný	trestný	k2eAgInSc4d1
čin	čin	k1gInSc4

Table 3: The document in vertical format with morphological annotation (after disambiguation)

Part of Speech	Count
k1 – noun	12884
k2 – adjective	4634
k3 – pronoun	2252
k4 – numeral	1028
k5 – verb	4504
k6 – adverb	933
k7 – preposition	3600
k8 – conjunction	3764

Table 4: Frequencies of part of speech in the document

3. Noun Groups

For the recognition of the noun groups we have used the partial syntactic analyzer for Czech DIS/VADIS (Žáčková, 2002) at first. Unfortunately, DIS/VADIS presently does not contain rules which can recognize genitival and coordinate structures because during the development of DIS/VADIS these rules were found too erroneous (overgenerating) when applied to an unrestricted text. However, there are plenty of such structures in the law texts and overgenerating is not a problem here because the results will be checked manually.

Moreover, the partial syntactic analyzer DIS/VADIS has one more disadvantage: it is written in Prolog which implies that the recognition process is rather slow. Therefore we have rewritten the rules for noun groups to Perl 5 regular expressions (which have non-trivial backtracking capabilities) and added the rules for genitival and coordinate structures and some adverbials common to the law texts which also were not recognized by DIS/VADIS (e.g. *zvlášť* (exceedingly), *zjevně* (evidently) etc.).

For each noun group found in the law texts we determine its:

Příprava	<l>příprava <c>k1gFnSc1 (preparation)
k	<l>k <c>k7c3 (to)
trestnému	<l>trestný <c>k2eAgMnSc3d1 <c>k2eAgInSc3d1 <c>k2eAgNnSc3d1 (criminal)
činu	<l>čin <c>k1gInSc3 <c>k1gInSc6 <c>k1gInSc2 <l>čina <c>k1gFnSc4 (act)
je	<l>být <c>k5eAaImIp3nSrDaI <l>on <c>k3p3gMnPc4xP <c>k3p3gInPc4xP <c>k3p3gNnSc4xP <c>k3p3gNnPc4xP <c>k3p3gFnPc4xP <l>je <c>k0 (is)
trestná	<l>trestný <c>k2eAgFnSc1d1 <c>k2eAgFnSc5d1 <c>k2eAgNnPc1d1 <c>k2eAgNnPc4d1 <c>k2eAgNnPc5d1 (criminal)

Table 2: Output of the morphological analyser Ajka

1. base form (nominative singular),
2. head
3. for nouns in genitive groups also their part.

For example for the noun group *dalším páchání trestné činnosti (subsequent commission of criminal activity, dative)* we get:

1. *další páchání trestné činnosti*
2. *páchání*
3. *další páchání*

We can recognize 8,594 noun groups counting repeating occurencies, 3,992 different noun groups. Table 5 lists several most frequent noun groups with the respective number of occurrences in the pilot data (since there are some conceptual problems with finding the correct English equivalent terms we do not offer them here).
The noun groups was analyzed and the respective 'base' of each noun group was derived. Due to the inflectional feature of Czech this cannot be done by simple lemmatization of all words in a noun group. The automatic transformation algorithm works in following steps:

- find dependences between parts (words of subgroups) of a noun group,
- locate the root – key word,
- identify matching noun group pattern,
- generate the correct word forms with matching gramatical categories.

The result of this algoritm are base forms of noun groups and they will appear as headwords in the final electronic dictionary. The most frequent base forms with respective number of occurrences in the pilot data are listed in Table 6.

Noun Group	Count
odnětím svobody	492
peněžitým trestem	139
jeden rok	123
trestný čin	79
odnětí svobody	76
účinnosti dne	65
zákazem činnosti	64
trestného činu	58
velkého rozsahu	49
závažný následek	47
zvlášť závažný následek	46
(jiné) majetkové hodnoty	46
těžkou újmu	44
značný prospěch	40
jiný zvlášť závažný následek	40
výjimečným trestem	39
organizované skupiny	39
člen organizované skupiny	39
značnou škodu	38

Table 5: The most frequent noun groups

Table 7 presents the most frequent part-of-speech patterns of the recognized noun groups. There are two counts in the table, 'Count Tokens' is the total number of occurrences of the respective pattern in the pilot data, 'Count Types' is the number of different noun groups matching such pattern.

4. Verb List

Though law terms typically consist of the nouns, noun groups and other nominal constructions we also have paid attention to the verbs found in the whole database of the 50,000 law documents. The reason for this comes from the fact that verbs on one hand do not display strictly terminological nature but on the other they are relational elements linking the terminological nouns and noun groups together. This can be captured by the surface and deep verb valency frames (Horák and Hlaváčková, 2005) of the verbs occuring in the law documents. We are not aware of any attempt to

Part of Speech Patterns	Count Tokens	Count Types
k2 – k1gI	1588	344
k2 – k1gF	1130	365
k1gN – k1gF	765	96
k2 – k1gN	478	213
k1gI – k1gN	204	57
k1gN – k1gI	203	80
k1gI – k1gF	195	67
k2 – k1gM	176	71
k2 – k2 – k1gF	163	65
k1gF – k1gI	162	48

Table 7: The most frequent POS patterns

Noun Group	Count
odnětí svobody	568
trestný čin	228
peněžitý trest	152
jeden rok	123
zákaz činnosti	81
trest odnětí svobody	69
účinnost dne	65
(jiná) majetková hodnota	65
velký rozsah	64
těžká újma	58
výjimečný trest	51
organizovaná skupina	49
závažný následek	47
zvlášť závažný následek	46
veřejný činitel	46
značný prospěch	40
jiný zvlášť závažný následek	40
značná škoda	39
člen organizované skupiny	39

Table 6: The most frequent terms

describe the valency frames of the verbs coming from law texts. Presently the verb list comprises 15,110 items, particularly 10,190 infinitives and 4,920 participles (which are mostly the passive ones). The list has been processed by the morphological analyzer Ajka (Sedláček, 2005) as a result we have obtained the list of 914 items that were not recognized by Ajka morphological tool. The structure of this list shows that at least three types of the non-recognized items can be observed:

1. erroneous forms caused by typing errors, they can be corrected, e. g. *cítit (feel),*

2. the verbs that Ajka does not know, i. e. the ones that do not appear in the Ajka's list of stems. Typically, they display a terminological character and they should be added to the Ajka's stem list, e. g. *derogovat (derogate)*. They will enrich the list of (Czech) stems and their law meanings constitute a terminological subset of verbs,

3. erroneous forms that cannot be corrected without correcting the whole paragraph of a law document (we do not touch them).

The next step is to add the non-recognized verbs to Ajka's list of the verb stems and to make an intersection with our existing database Verbalex (Horák and Hlaváčková, 2005) containing presently about 11,306 (general) Czech verbs.

5. Context patterns in law texts

We take the position that the decisive information about the semantics of the law terms comes from the contexts in which they occur. There are two ways how to approach this:

- To use statistical techniques by means of which we obtain the interesting contexts – they can be sorted and the semantic clusters they create can be built. The limitation here is that the data from the law texts are not large enough and in some cases we do not get enough contexts to make the necessary generalizations.

- To explore the valency frames in the law texts and find the semantic roles that are typical for the verbs in the law texts. We already have done this for approximately 11 000 of (general) Czech verbs and the result is that we learn enough not only about the verb meanings but also about arguments constituting the argument-predicate structure of the 'law' verbs.

We expect that the inventory of the semantic roles for 'law' verbs will reasonably differ from the 'general'

verbs and, on the other hand, that there will be interesting polysemy which we capture by means of semantic roles occurring in the found valency frames of the 'law' verbs. The two approaches, obviously, can be combined.
To show how we understand valency verb frames and the corresponding semantic roles we offer the example with the two following verbs:

1. *uložit trest někomu (to condemn sb to a sentence)* The meaning of this verb can be described by the following frame:
   AG(judge:1)[1] – uložit – PAT(person:1)[3] ACT(sentence:1)[4]

2. *obvinit z trestného činu koho (accuse sb of criminal act)* The meaning of this verb can be captured by the following frame: AG(public prosecutor:1)[1] – obvinit – PAT(person:1)[4] ACT(act:2)[z2]

To explain briefly the notation used: for the semantic roles we use labels like AG(judge:1), which say that the agent of the verb *uložit (condemn)* has to be a judge, the second role can be any person and the third one is the ACT, i.e. a sentence. Moreover, the labels used for the roles are closely linked to Princeton WordNet literals and they represent nodes in this semantic network which yields relevantinformation about their senses. The numbers following the roles express morphological cases that have to be indicated in Czech.
The frames adduced here are very similar to the frames as they presently exist in our verb frame database Verbalex mentioned above. This means that the effort put into its building can be exploited also in the area of the law texts. The more important thing, however, is that the valency frames capture noun and prepositional groups obtained via morphological and syntactic analysis mentined above and tell us what is their meaning. In other words, this knowledge allows us to find out what entities are denoted by noun and prepositional groups in law text and on this ground to build an ontology for the law domain. Then it can be compared with the already existing law ontologies such as the one built within the LOIS (Lexical Ontologies for Legal Information Society) project[1]. In this project the ontology is built in the WordNet fashion. It can be expected that the ontology exploiting semantic roles in valency frames should be closer to law texts in their natural form. Thus we can conclude that building valency frames of verbs occuring in law text is one of the important tasks set in this project.

6. Conclusion

We have presented the preliminary results of the computational analysis of Czech law documents, or more precisely, their selected samples. On one hand we have used the already existing tools such as Ajka or DIS/VADIS, on the other hand we have modified them respectively for the purpose of the present task. As a result we can enrich them with regard to the law language but, more importantly, we have obtained basic knowledge about the grammatical structure of the law texts (law terminology) and in this way we are prepared to continue our exploration of the contexts in which law terms occur in the law documents.
The knowledge of such contexts is a necessary condition for a deeper understanding of how law terminology works and how it can be made more consistent. As an application we intend to obtain the basic rules for intelligent searching law documents. A tool based on such rules can serve to judges, attorneys and experts in creating new law documents. In other words, the relevant output of this work thus will be an electronic dictionary of law terms.

Acknowledgements

This work has been partly supported by the Academy of Sciences of the Czech Republic under the projects 407/07/0679 and by the Ministry of Education of the Czech Republic within the Centre of basic research LC536.

7. References

A. Horák and D. Hlaváčková. 2005. VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages, Third International Seminar*, pages 107–115, Bratislava. VEDA.

Pavel Rychlý. 2000. *Corpus managers and their effective implementation*. Phd thesis, Faculty of Informatics, Masaryk University.

B. M. Schulze and O. Christ, 1996. *The CQP User's Manual*.

Radek Sedláček. 2005. *Morphemic Analyser for Czech*. Ph.D. thesis, Masaryk University.

Pavel Šmerk. 2007. *Towards Morphological Disambiguation of Czech.* Ph.D. thesis proposals, Faculty of Informatics, Masaryk University.

Eva Žáčková. 2002. *Partial syntactic analysis of Czech*. Phd thesis, Faculty of Informatics, Masaryk University.

[1]see http://nlpweb.kaist.ac.kr/gwc/pdf2006/50.pdf and also http://www.ittig.cnr.it/Ricerca/materiali/lois/WhatIsLOIS.htm