Morphological Analysis of Law Texts

Karel Pala, Pavel Rychlý, and Pavel Šmerk

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

Abstract. In the paper we explore the morphology of the Czech law texts including Constitution, acts, public notices and court judgements which form a huge textual database. As many texts from small domains, the used language is partially restricted and in relevant aspects also different from general Czech.
The paper presents first results of the morphological analysis of Czech law texts and their conversion to the specific formats. Partly, the partial syntactic analysis has been performed as well.

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 start with a legal database Lexis which presently includes approx. 50,000 Czech law documents ranging from the beginning of Czechoslovakia in 1918 to present days. It also includes court judgements, main representative law textbooks and law reports. All the texts exist in electronic form.

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.

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 morphological 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 lemmatizations are transformed into a special format which is described below.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007, pp. 21–26, 2007.

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

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

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. 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 [1] and Manatee [2].

In the next step, we processed the vertical file with the morphological analyser Ajka [3]. 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. In the following 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.

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)

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 [4] to disambiguate such word forms. The output of the Desamb tool contains only the most probable lemma/tag for each word form. Table 2 contains output of Desamb for the input text above.

Table 2. The document in vertical format with morphological annotation (after disambiguation)

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

The annotated version of the document contains 2,560 different lemmas. Frequencies of each part of speech are in Table 3.

Table 3. Frequencies of part of speech in the document

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

3 Noun Groups

For the recognition of the noun groups we have used the partial syntactic analyzer for Czech DIS/VADIS [5] 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 nontrivial 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:

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 4 lists several most frequent noun groups (since there are problems with finding the correct English equivalent terms we do not offer them here). Table 5 presents the most frequent part-of-speech patterns of the recognized noun groups.

Table 4. The most frequent noun groups

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

Table 5. The most frequent POS patterns

Part of Speech Patterns	Count
k2 k1gI	1526
k2 k1gF	1127
k1gN k1gF	769
k2 k1gN	469
k1gI k1gN	210
k1gN k1gI	203
k1gI k1gF	193
k1gF k1gI	177
k1gF k1gN	171
k1gF k1gF	164

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 [6] of the verbs occuring in the law documents. We are not aware of any attempt to 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 [3] as a result we have obtained the list of 914 items that were not recognized by Ajka 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. *cítít (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.

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 [6] containing presently about 11,306 (general) Czech verbs.

5 Conclusion

We have presented the preliminary results of the computational analysis of Czech law documents, or more precisely, their samples. On one hand we have used the already existing tools such as Ajka or DIS/VADIS, on the other hand we have modified respectively them 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 hope 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.

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.

References

1. Schulze, B.M., Christ, O.: The CQP User's Manual. (1996).
2. Rychlý, P.: Corpus Managers and their Effective Implementation. Ph.D. thesis, Faculty of Informatics, Masaryk University (2000).
3. Sedláček, R.: Morphemic Analyser for Czech. PhD thesis, Masaryk University (2005).
4. Šmerk, P.: Towards Morphological Disambiguation of Czech. Ph.D. thesis proposals, Faculty of Informatics, Masaryk University (2007).
5. Žáčková, E.: Partial Syntactic Analysis of Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University (2002).
6. Horák, A., Hlaváčková, D.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Computer Treatment of Slavic and East European Languages, Third International Seminar, Bratislava, VEDA (2005) 107–115.