Legal Electronic Dictionary for Czech

František Cvrček, Karel Pala, Pavel Rychlý

Institute of State and Law, Institute of State and Law,
Czech Academy of Sciences, Prague, Faculty of Informatics Masaryk University, Brno
Czech Republic
f.cvrcek@worldonline.cz, pala@fi.muni.cz, rychly@fi.muni.cz

Abstract

In the paper the results of the project of Czech Legal Electronic dictionary (PES) are presented. During the 4 year project the large legal terminological dictionary of Czech was created in the form of the electronic lexical database enriched with a hierarchical ontology of legal terms. It contains approx. 10,000 entries – legal terms together with their ontological relations and hypertext references. In the second part of the project the web interface based on the platform DEBII has been designed and implemented that allows users to browse and search effectively the database. At the same time the Czech Dictionary of Legal Terms will be generated from the database and later printed as a book. Inter-annotator's agreement in manual selection of legal terms was high – approx. 95 %.

Keywords: legal dictionary, electronic database, Web interface

1. Introduction

In this paper we describe current work on creating a Czech legal electronic dictionary (Czech title is – Právnický elektronický slovník, further referred to as PES). The data, on which PES is based, have been prepared in the Institute of State and Law, Academy of Sciences of Czech Republic. It is a large database of legal texts containing approx. 180,000 documents comprising legislation, judicature, textbooks and commentaries. This database represents the Czech legal system since 1918, i.e. since the founding of Czechoslovakia as an independent state. The creation and analysis of these resources has started about 1986. The results to date are summarized also in the book Legal Informatics (cek, 2011).
The design of the Dictionary is based on the following assumptions:

1. The number of terms contained in the texts of existing valid Czech laws is only a fraction of terms that are commonly used for their interpretation. Eg. the Czech Criminal Code contains 700 basic legal terms and the textbook based on this law contains only in the basic part over 2,000 legal terms. The intersection between the legal terms from textbooks and legal terms from the statute under discussion is about 10%. This is also the case for commentaries and other legal texts used in the practical application of legal regulations.

2. The legal language is, from the point of view of the occurrence of legal terms, internally divided: there are sublanguages for the individual legal branches (their number is about 15-18) and also for the language of the primary legal regulations, secondary regulations, judicature, etc. Thus legal language is a specialized language for experts. The same legal term may change its meaning from one legal sublanguage to another.

3. The meaning of legal terms is determined by the context and structure to which they belong. Most legal texts are strongly dependent on what may be called linearity of the text, where it is assumed that a recipient will read the text sequentially, i.e. as words following each other, provided that he/she will understand their meaning from the context in which they occur. However, the context necessary for an interpretation of the text need not be close context only.

4. In the legal tradition from the time of glossarists it has been usual to segment a legal text into sections (text paragraphs containing internal content headings). Segmentation of the text, i.e. finding its structure, is highly important for its understanding and affiliation of the normative legal text to a certain section can characterize its meaning. For example, the definition of murder in the particular section of the Criminal Code contains the word *murder* only in the section and not in a normative text as such. Or structure of the sections (we will further speak about the titles of the sections) may determine whether the period, which is mentioned in the normative text is an objective or subjective period, etc.

5. Many similar legal dictionaries exist, e.g. for French (Cornu, 1987), English (Garner, 2009; Duhaime, 2011; Dean, 2011) and other languages. In our view, the main difference between PES and other legal dictionaries consists in systematic emphasis on processing contexts in which legal terms occur and relations between them.

2. Czech legal electronic dictionary – the project PES

The main goal of the PES project (GaČr 2009-11) is to use contemporary information technologies and linguistic techniques to create not only a lexical database of legal terminology involving the actual meaning of legal terms in plain definitions, in their context and structure, but, as we said above, also a Czech legal electronic dictionary.

- The PES project integrates three different approaches:

1. building of the classic explanatory dictionary of legal terms,
2. contextual dictionary of legal terms, and
3. structural links between the legal terms.

The underlying task is to build software tools that will allow the users to access and browse the dictionary through a web interface.

Basic entries have been produced by the manual selection of the headwords from textbooks and commentaries respecting the fact that documents should be informatively exhausted. This assumes that a person processing the text is an expert in the field, who is able to recognize terms and include them into the structure of the document. The main emphasis is on structural integration, a procedure that respects the author of the document. Often, the author has to explicate an implicit structure, which can be a complicated problem. Each entry is assigned a location in the hierarchical structure up to the ultimate integration into the particular legal branch.

If the main entry is composed of several words, the person processing the text must decide whether it is from a legal point of view one term or more terms. For example, in Czech criminal law, there is a term *especially serious recidivist*, which, unlike the term *recidivist* does not exist in the criminal law, but in textbooks and judicature it is regularly used, usually in the sense particularly serious recidivist. If a term is compound, the expert processing it has to decompose it as well.

The Dictionary also describes the relations of the association, synonymy and antonymy. These relationships are relative and have meaning only in the link to a current structure of the particular document. Unfortunately, the authors may differ in what they understand as synonymy, association or antonymy. Definition of the entry refers to the part of text that contains the term and explains it or adds a normative value to it. In many cases, it can be a mere reference to a particular law.

The definitions are classified according to their realization. For example, we can have legal definitions, theoretical definitions, item enumerations, lists of examples, references to a paragraph of the law, etc.

If the definition consists of items related by coordinated conjunctions, they will be marked to distinguish them from the items connected by disjunction. Interpretation of the connection of items has to be done manually, because the words *and* or *or* may not mean the conjunction or disjunction in the legal sense.

- A special part of the Dictionary is a collection of structured sections of legal regulations. The section is then understood as a basic entry. We have processed the Czech legislation in a standardized form, so the detection of sections and their hierarchical relations was performed automatically. If section's meaning depends on the context the context is automatically filled in from the higher level sections.

- The last part of the Dictionary is a collection of legal terms from legal regulations based on the automated analysis of frames consisting of the part of speech tags, such as AAN – *organizovaná zločinecká skupina (organized criminal group)*. We also dealt with automatic identification of legal terms in Czech law texts (Pala et al., 2010). Accuracy of the automatic detection of most frequent patterns is approx. 90%, the rest of them has to be classified manually. Within the PES project we then display the most frequent contexts obtained in this way from legal regulations.

3. PES Dictionary Structure

The structure of the PES has the following form:

- number of the entry – it expresses an order of the entry in the particular document.

- legal branch – e.g. Criminal law, Constitutional law, Civil law, Substantive law, Roman law, historical legal terms, etc.

- entry – the basic unit of the dictionary which comes from the marked part of the text from a textbook, or it is a legal definition from law instructions or section titles.

- type of the definition – theoretical definition, legal definition, enumerative definition, definition via reference to the particular law or judicature, definition by example, definition of the legal principle, etc.

- English equivalent of the entry – if there is an official translation of the document it is used, in other cases, the translations from standard legal Czech-English dictionaries are used.

- synonyms, associated terms or antonyms – they are offered by the experts but they must be based on the text of the definition occurring in the given text. In this sense these items have to be considered relative.

- internal elements (inelements)– they are variables that represent decomposition of the complex entry to the atomic elements. Atomic entries do not contain inelements.

- section titles – they describe structural placement of the entry in a particular text. In other words, they can be understood as nodes in the tree of legal terms.

- definition – it is understood as an explication of the entry meaning. It is either a reference to the particular law text – then it can be recognized automatically, or in the case of the textbooks or commentaries it is given by an expert.

- type of the resource – information that a given document is a textbook, commentary or law.

- definition member – it is related to a decomposition of the definition to the atomic legal terms. In this sense they are similar to the inelements.

- constituents and sorts – they determine relations between definition members. These relations can be either conjunctions (constituents) or disjunctions (sorts). The sorts then have a form of the complete enumerations or demonstrative ones.

- authors – for each entry, of section title, definition, inelement, member, constituent and sort the author is given together with the reference to the respective page of the document or paragraph of the law text.

- notes – additional information provided by the document authors to the given definition (reference to a related literature, related laws, etc). Also the author of the note is usually given.

The entries in dictionary are ordered in the tree of section titles and branches to which legal terms belong. The tree can be displayed from the root (reference to the tree of section titles) or as a subtree when we click on the part of the found path.

3.1. PES Dictionary as a Base for the Analysis of Legal Language

The data in the PES database provide a very interesting material for various kinds of analyses.

- Terminological decomposition of the individual legal branches together with their mutual relations makes it possible to perform a number of analyses. The first attempts have brought interesting results that make many traditional law ideas doubtful.

- If we compare entries and structures representing, on the one hand, a law and, on the other, textbooks for the basic legal branches we come to the following findings: textbooks contain 10 times more legal terms than the corresponding laws. The intersection of the laws and textbooks is slightly below 10 %. Legal definitions in Czech laws represent 5 – 7 % legal terms from these laws. The depth of the hierarchical tree of section titles is half in comparison with depth of the textbook tree. Though in the continental culture we pretend that the basic source of the legal system is law it appears that for its interpretation we need a terminological apparatus ten times larger than the one on the level of texbooks. For the judicature and the commentaries these data are still in processing.

- The following table brings an interesting comparison, which describes the numbers (in %) of various inelements, entries and section titles in relation to whole number of the terms from the branch of the Criminal law and Roman law.

  From the Table 1 it follows that the largest percent of the entries from the whole number of the terms of given legal branch is defined in Roman law (61 %), then in the Substantive Criminal law (53 %) and the smallest in the Criminal law (41 %). The smallest number of the Inelements is contained in Roman law (41 %) and the largest one we find in the Criminal law. The highest percent of section titles is in the Roman law (25 %) and the lowest in the Criminal law (8 %). At the same the Roman law contains the highest percent of the defined Inelements (17 %) and the lowest percent (5 %) is found in the Criminal law.

  It follows that the Roman law is, from the terminological point of view, the most elaborated legal branch from all presented documents, both with respect to the number of the defined entries, defined Inelements and section titles.

	I	H	R	S	R and H	I and H
TPH	51	53	19	100	86	12
TPP	53	48	17	100	70	9
RIM	41	61	25	100	75	17
TZ	56	41	8	100	30	5

Table 1: Comparison of different counts (in %) from Criminal law and Roman law
legend: I – inelements, H – entries, R – section titles, S – Suma Term;
TPH – textbook of the Substantive Criminal law, TPP – textbook of the Procedural Criminal law, RIM – textbook of the Roman law, TZ – Criminal law. Term is a complete number of the terms for a given document (100 %).

- The section titles represent in the PES dictionary a hierarchical structure with one root and up to 10 levels. The entries can be understood as its leaves (or nodes). The numbers of the nodes on the individual tree levels can be seen in the Table 2.

	K1	N1	N2	N3	N4	N5
TPH	0.05	0.89	4.55	22.2	38.21	23.24
RIM	0.23	0.23	0.9	3.15	9.68	17.12

	N6	N7	N8	N9	N10
TPH	7.38	2.73	0.55	0.25	0
RIM	24.1	21.17	17.12	5.63	0.68

Table 2: Numbers of nodes on individual tree levels
legend: TPH is a tree from the textbook of Substantive Criminal law and RIM is a tree from the Roman law textbook;
K1 is a root of the tree, N1 – N10 are numbers of the nodes in percents on the respective levels

  TPH is a tree from the textbook of Substantive Criminal law and RIM is a tree from the Roman law textbook. K1 is a root of the tree, N1 – N10 are numbers of the nodes in percents on the respective levels. Table 2 also shows the distribution of the nodes in both trees. While for TPH we observe the highest percent of the nodes on levels N3 – N6 (more than 5 %), for RIM the highest percent of the nodes on the levels N4 – N9 (above 5 %).

  This is, of course, just an elementary example, which should demonstrate, that the analysis of the structures can offer interesting results. If it appears, for example, that our hypotheses can be verified and some legal branches will demonstrate a higher measure of complexity this should have consequence for the conception of legal studies.

- Then we can ask different questions, for instance, which variable or their combination can provide a set of the first 10 terms that characterize a given legal branch in the best way? As the best choice can be considered such subset of the legal terms, which differentiates a given legal branch from others (information weight) and at the same time the experts will agree on selecting it as representing this legal branch best.

  These results are interesting since they show that the principal legal terms are not comprised in the entries, i.e. they are not defined at all but they come from the decomposition of the entries or from the structure of section titles. Thus it seems that we can understand the role of the basic legal terms rather through the structural position of the entry (external and internal structure) than via the simple relation entry – definition.

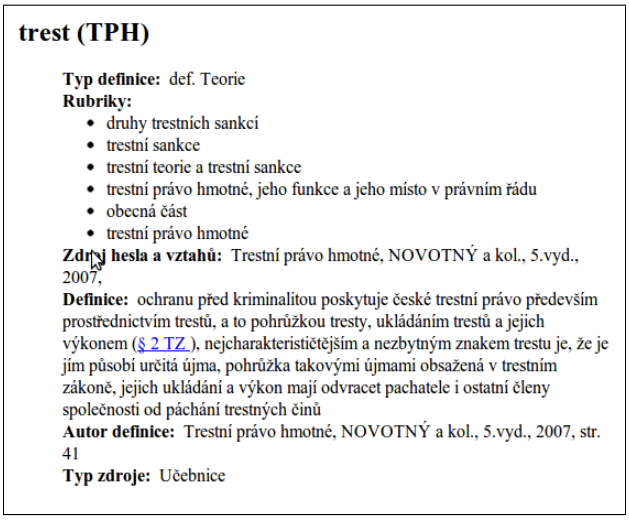
  It is obvious that the usual search in the electronic dictionaries based on the fulltext search in the entries and definitions, especially on the assumption of the large size of the dictionaries with more than 100,000 entries, has to be regarded as quite primitive. Our experience with searching in standard legal information systems tells us that the simple searching via the section titles improves the effectivity of searching in large legal databases up to 10 times. Thus it can be justifiably expected, that the PES dictionary would enable the creation of the more sophisticated searching functions based on the analysis of the both internal and external structures of the primary entries.

4. Software tools for PES

The Dictionary data created by law lexicographers from the Institute of State and Law have been edited first in a spreadsheet tables. The entries are grouped into separate files depending on the individual legal branches. Presentation of the Dictionary is provided by an appropriate web application. Such solution has many advantages, the most important is that users do not need to install any software on their computers and new versions of the application are available immediately for all users.

The presentation application is based on DEBII platform (Horák et al., 2006). It is a platform for dictionary editing and browsing, it uses XML as a main data format and a powerful XML database engine for storing and retrieving the data. The initial application for PES is a server only system, which consists of several scripts providing visual content in the form of HTML pages. There are query forms and displaying list and individual entry pages, all pages are linked together. Current development of the application leads to a client server system, which is more dynamic and more user friendly.

For the transition of the Dictionary from original tables into XML format suitable for importing it into DEB XML database an automatic system was developed. This system processes the source data in several steps, each step makes a simple transformation or adds some bits of information to the data. The initial steps work on data formats with one XML file as a result. The following steps add unique ID numbers to entries and links from one entry to another, add lemmatization for faster retrieving and create separate ontology tree of all entries in the Dictionary.

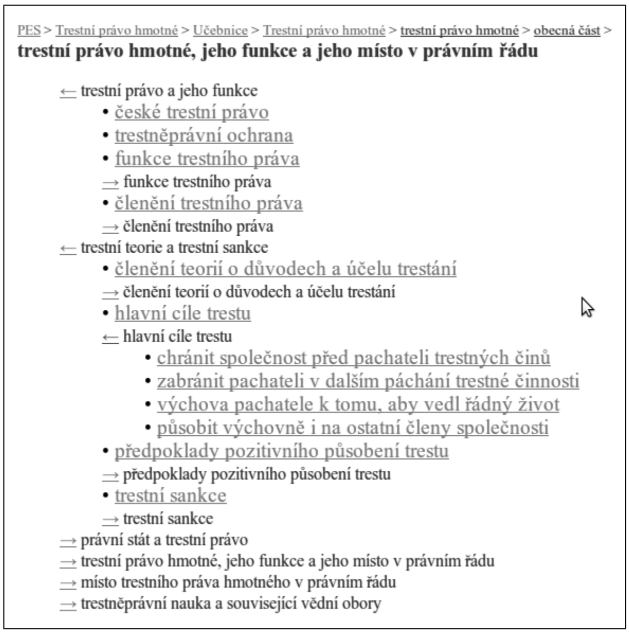


trest (TPH)

Typ definice: def. Teorie
Rubriky:
- druhy trestních sankcí
- trestní sankce
- trestní teorie a trestní sankce
- trestní právo hmotné, jeho funkce a jeho místo v právním řádu
- obecná část
- trestní právo hmotné

Zdroj hesla a vztahů: Trestní právo hmotné, NOVOTNÝ a kol., 5.vyd., 2007,
Definice: ochranu před kriminalitou poskytuje české trestní právo především prostřednictvím trestů, a to pohrůžkou tresty, ukládáním trestů a jejich výkonem (§ 2 TZ), nejcharakterističtějším a nezbytným znakem trestu je, že je jím působí určitá újma, pohrůžka takovými újmami obsažená v trestním zákoně, jejich ukládání a výkon mají odvracet pachatele i ostatní členy společnosti od páchání trestných činů
Autor definice: Trestní právo hmotné, NOVOTNÝ a kol., 5.vyd., 2007, str. 41
Typ zdroje: Učebnice

Figure 1: Example of the entry *trest*.



PES > Trestní právo hmotné > Učebnice > Trestní právo hmotné > trestní právo hmotné > obecná část >
trestní právo hmotné, jeho funkce a jeho místo v právním řádu

← trestní právo a jeho funkce
- české trestní právo
- trestněprávní ochrana
- funkce trestního práva

→ funkce trestního práva
- členění trestního práva

→ členění trestního práva

← trestní teorie a trestní sankce
- členění teorií o důvodech a účelu trestání

→ členění teorií o důvodech a účelu trestání
- hlavní cíle trestu

← hlavní cíle trestu
  - chránit společnost před pachateli trestných činů
  - zabránit pachateli v dalším páchání trestné činnosti
  - výchova pachatele k tomu, aby vedl řádný život
  - působit výchovně i na ostatní členy společnosti
- předpoklady pozitivního působení trestu

→ předpoklady pozitivního působení trestu
- trestní sankce

→ trestní sankce

→ právní stát a trestní právo
→ trestní právo hmotné, jeho funkce a jeho místo v právním řádu
→ místo trestního práva hmotného v právním řádu
→ trestněprávní nauka a související vědní obory

Figure 2: Example of the partially expanded ontology.

4.1. DEB Platform

DEB Platform is the freely accessible lexicographical platform (`http://deb.fi.muni.cz/`) for browsing, building or editing of electronic dictionaries. DEB uses the architectures client-server. The data are stored in the XML database (Oracle Berkeley DB XML, or Sedna). For servlets (server parts of the applications which serve to the database etc.) DEB uses the programming language Ruby. The queries for XML databases exploit languages Xpath and XQuery.

5. Results

The following results have been obtained so far:

1. legal lexical database containing approx. 10,000 entries – legal terms together with their ontological relations and hypertext references – they have been obtained from the legal texts manually.

2. Then we have extracted 20,000 terms from legal regulations – these have been obtained automatically – we have been using as a definition their reference to a paragraph, in which they occur. This is the second part of the built legal database.

3. The web interface based on the platform DEBII has been designed and implemented that allows users to browse and search the database. At the same time the Czech Dictionary of Legal Terms will be generated from the database and printed later as a book.

4. The number of the single word legal terms is presently about 700, they come from legal regulations and they had to be selected manually. Another 700 single word terms come from judicature.

5. Inter-annotator's agreement reached in manual selection of terms was high – approx. 95 %.

6. Conclusions

Processing of the structures in the individual documents is partly automatic – in the case of the section titles of legal regulations, and partly manual – in the case of textbooks. This division is one of basic points of our approach. The Dictionary outlined above also describes a given entry from various aspects, i.e. how authors of textbooks, law, judicature work with it in their texts.

The number of basic entries for the Czech legal code is estimated approx. at 10,000 items. This is in a good agreement with Dean's dictionary (Dean, 2011). The average textbook for a branch of law contains 2,000 to 5,000 terms. Number of contextual terms obtained from the legal regulations will be hundreds of thousands, but these can be generated automatically from the subordinate legal regulations mainly. Number of entries obtained by the analysis of the section titles is estimated for the current core of legal regulations as 20,000 items.

7. Acknowledgements

This work has been supported by the Academy of Sciences of Czech Republic under the project Agency under the project GA 406/12/1514. Part of this research was also related to the Lindat-Clarin project supported by the Czech Ministry of Education.

8. References

F. Cvrček. 2011. *Právní informatika (Legal Informatics)*. Ústav státu a práva AV ČR (The Institute of State and Law CAS), nakl. A. Čeněk (Publishing House A. Čeněk), Prague.

Gerard Cornu. 1987. *Vocabulaire Juridique*. French and European Publications, Incorporated.

Dean. 2011. Dean's law dictionary. http://www.deanslawdictionary.com/.

Duhaime. 2011. Duhaime's legal dictionary. http://www-.duhaime.org/LegalDictionary.aspx.

Bryan A. Garner. 2009. *Black's Law Dictionary*. West Law School, 9th edition.

A. Horák, K. Pala, A. Rambousek, and P. Rychlý. 2006. New clients for dictionary writing on the DEB platform. In *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems*, pages 17–23, Torino, Italy. Lexical Computing Ltd.,.

K. Pala, P. Rychlỳ, and P. Šmerk. 2010. Automatic identification of legal terms in czech law texts. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 83–94. Springer Verlag, Berlin, Heidelberg.