Czech Legal Text Treebank 1.0

Vincent Kríž, Barbora Hladká, Zdeňka Urešová
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{kriz, hladka, userova}@ufal.mff.cuni.cz

Abstract

We introduce a new member of the family of Prague dependency treebanks. The Czech Legal Text Treebank 1.0 is a morphologically and syntactically annotated corpus of 1,128 sentences. The treebank contains texts from the legal domain, namely the documents from the Collection of Laws of the Czech Republic. Legal texts differ from other domains in several language phenomena influenced by rather high frequency of very long sentences. A manual annotation of such sentences presents a new challenge. We describe a strategy and tools for this task. The resulting treebank can be explored in various ways. It can be downloaded from the LINDAT/CLARIN repository and viewed locally using the TrEd editor or it can be accessed on-line using the KonText and TreeQuery tools.

Keywords: annotated corpus, legal domain, parsing

1. Introduction

In our work, we develop approaches and systems for detecting semantic relations from unstructured texts. We see this task as one of the most important component for search engines which could become more sophisticated and user-friendly for querying textual documents.

We developed the RExtractor[1] system for detecting semantic relations from unstructured texts (Kríž et al., 2014; Kríž and Hladká, 2015). The system extracts a knowledge base from raw unstructured texts. The knowledge base is a set of entities and their relations represented in an ontological framework.

The RExtractor system implements an extraction pipeline which processes input texts by linguistically-aware tools and extracts entities and relations using queries over dependency trees. The language used for testing RExtractor is Czech and the legal domain was chosen to be explored in detail.

We have surveyed available syntactically annotated corpora. Only a very few contain some texts from the legal domain - e.g., some of the smaller corpora from the Universal Dependencies set.[2]

A syntactic parsing used in the RExtractor pipeline is of a crucial importance for the extraction. Because of lack of any Czech gold legal-domain data, we have used the MST parser (McDonald et al., 2005) trained on the Prague Dependency Treebank (PDT, (Bejček et al., 2013)), i.e., on newspaper texts. We thus had to create a gold data set from the legal domain, in order to get at least a rough idea about the performance of the parser on a domain that is different from the domain of the parser's original training data; this has resulted in the Czech Legal Text Treebank (CLTT). In total, 1,128 sentences from the Collection of Laws of the Czech Republic were annotated morphologically and syntactically in accordance with the Prague Dependency Treebank annotation framework.

In addition, we have manually annotated entities and their relations in CLTT in order to evaluate RExtractor as a whole. We measured the RExtractor performance on a part of CLTT, namely on 762 sentences. The system achieved precision of 80.6% and recall of 63.2%. We identified three sources of errors: (i) incorrect dependency tree (59.7%), (ii) missing or incorrectly formulated query (38.3%), (iii) missing or incorrectly recognized entity (2.1%). We can see that errors are mainly caused by the insufficient quality of syntactic parsing. It confirms crucial importance of improving the quality of the RExtractor parsing component by using in-domain data (i.e., CLTT) for training (and cross-validate it on the same domain).

The remainder of this paper is organized as follows. Section 2. presents a brief description of CLTT. We focus on studying and comparing various language phenomena present in CLTT and PDT in Section 3. Details on CLTT, namely on the annotation layers and the annotation process are provided in Section 4. Finally, Section 5. provides information about accessing and getting CLTT and Section 6. presents our plans for future work.

2. Czech Legal Text Treebank 1.0

The CLTT 1.0 consists of 35,058 tokens in 1,128 morphologically and syntactically annotated sentences.

2.1. Text Sources

The CLTT contains two legal documents: (1) The Accounting Act (563/1991 Coll., as amended) and (2) Decree on Double-entry Accounting for undertakers (500/2002 Coll., as amended). The selection was motivated by the goals determined in the Intelligent Library[3] project (Nečaský et al., 2013).

2.2. Annotation Layers

Dependency parsing of Czech legal texts fits the framework originally formulated in the Prague Dependency Treebank

[1] The system is available on-line: http://quest.ms.mff.cuni.cz:14280/

[2] http://universaldependencies.org/; 7 out of the 35 treebanks contain part described as belonging to the legal domain, and neither English nor Czech is among them.

[3] http://ufal.mff.cuni.cz/intlib

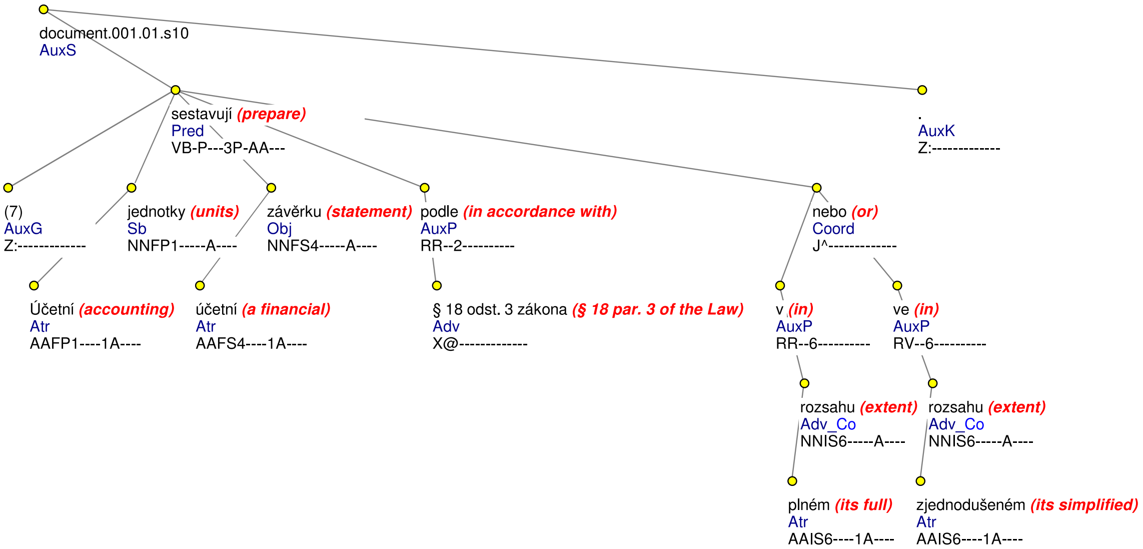


Figure 1: The dependency tree for Example 1.

project. The annotation of CLTT covers both the morphological[4] (m-layer) and analytical[5] (a-layer) layer. In addition, there is a non-annotation word layer (w-layer) representing a raw text segmented into documents, paragraphs, and tokens. A node of a tree expressing analytical annotation of a sentence is called the a-node.

The CLTT consists of 1,128 manually annotated dependency trees where each node corresponds to one token. The morphological annotation of each token was done automatically.[6] For illustration, consider Example 1 and its dependency tree visualized in Figure 1 where each a-node is accompanied by a morphological tag and an analytical function.

(1) (7) Účetní jednotky sestavují účetní závěrku
(7) Accounting units prepare a financial statement
podle §18 odst. 3 zákona v plném
in accordance with §18 par. 3 of the Law in its full
rozsahu nebo ve zjednodušeném rozsahu.
extent or in its simplified extent.

(7) Accounting units prepare a financial statement in accordance with §18 par. 3 of the Law in its full extent or in its simplified extent.

2.3. Data Format

We used the Prague Markup Language (PML, Pajas and Štěpánek (2006)) as a main data format. The PML is a generic XML-based data format designed for the representation of the rich linguistic annotation of text. Each of the annotation layers is represented by a single PML instance.

3. Studying CLTT

A legal text is something very different from ordinary speech. This is especially true of authoritative legal texts: those that create, modify, or terminate the rights and obligations of individuals or institutions. Lawyers often refer to them as operative or dispositive. Legal texts are specialized texts operating in legal settings. They should transmit legal norms to their recipients, therefore, they should be clear, explicit and precise. However, the style of legal texts is generally considered very difficult to read and understand.[7]

According to the theory of functional styles (as developed for Czech by the Prague School, primarily in Havránek's work (Havránek, 1932) and elaborated by many Czech scholars up to today, e.g., (Kořenský, 1989; Jelínek, 1995; Minářová et al., 2003)), the *function of the utterance* in communication is emphasized. This functional approach is based on goal-oriented language means and distinguishes several functional styles such as professional style, poetic style, colloquial style, etc. We are aware of the fact that the classification of the individual functional styles is a very complicated problem as mentioned e.g., in (Tiersma, 1999) or in (Gibbons, 2008).

However, having in mind the theoretical concepts of Czech functionally-oriented linguistics and general characteristics of the individual styles we tend to classify legal texts as texts belonging to the administrative-legal style (according to (Jelínek, 1996)) which is now earmarked as a unique functional style, standing next to other styles, such as professional, journalistic, literary or scientific. However, due to their specific function legal texts in many ways overlap with the professional style. Legal texts include very specific features related not only to vocabulary and syntax but also to various conventions and punctuation use. For example, impersonal style of legal texts understandably excludes the use of question marks and exclamation marks. On the other hand, we observe an extremely high usage of semicolon for purposes like enumeration, itemization and various types of listings.

For our purposes, the most important feature of legal texts is that they have a very specific syntactic structure with many peculiarities. We often encounter e.g., passive voice structures, impersonal constructions, non-finite and verbless clauses and conjunctive groups. Simple sentences are very rare. Typically, sentences are long and very complex. Punctuation plays a crucial role because legal texts usually include very complicated syntactic pat-

[4] http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/index.html

[5] http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html

[6] http://hdl.handle.net/11858/00-097C-0000-0001-48FE-9

[7] http://www.languageandlaw.org/LEGALTEXT.HTM

CORPUS	# OF WORDS	# OF SENTENCES	ASL
CNC	2,685,127,310	178,499,972	15.0
PDT 3.0	1,957,247	115,844	16.9
CLTT 1.0	35,085	1,128	31.0

Table 1: Average sentence length (ASL) of the Czech National Corpus (CNC, (Hnátková et al., 2014)), the Prague Dependency Treebank (PDT 3.0, (Bejček et al., 2013)), and the Czech Legal Text Treebank.

terns or long lists separated by semicolons.

The complexity of sentences present in CLTT is obvious even from such a simple measure like the average sentence length when compared to selected Czech corpora, see Table 1.

Despite the fact the legal texts should be clear, comprehensible and explicit we found them sometimes difficult to understand and annotate, because of high usage of syntactic condensation and unusual language patterns, significant tendency to prefer abstract expressions, nominalizations, chains of genitive expressions etc.

Table 2 documents a comparison of some selected language means as used in the CLTT corpus and the PDT 3.0 corpus (Mikulová et al., 2013). The PDT 3.0 corpus is a corpus of journalistic style and contains also genres annotation, see Table 3. The genres classification was originally created for the Prague Discourse Treebank 1.0 (Poláková et al., 2013) aiming to observe how the discourse relations function in different types (in the genre sense) of language (Poláková et al., 2014). In order to get a detailed picture, Table 2 includes all 19 genre categories (see the GENRE column) as classified in PDT 3.0 even though CLTT does not have genres classification at its disposal and it is considered to belong to a homogeneous "legal" genre.

Table 2 shows that legal texts are about 4.5 times "richer" in using a reflexive passive constructions while the use of periphrastic passive slightly prevails averaged across all genres in PDT – see the columns (1) and (2). The next columns (3), (4) and (5) document the expected dominance of chaining constructions with four, three and two genitives, respectively, in legal texts. The biggest difference (percentage ratio) is observed in constructions with two genitives; the CLTT legal texts use noun phrases with (at least) two genitives about 4.4 times more often than the PDT texts on average, with the ratio ranging from 3.06 (news) to more than 10 (interviews – not surprisingly, people do not use these genitive chains much when speaking).[8]

Our comparison did not confirm our assumption of a frequent use of the construction with deverbative nouns ending on *–ní*, *-tí* with genitive – see the column (6). Such a construction does not appear neither in the CLTT nor in the PDT texts very often. Surprisingly enough, we observe that the constructions with apposition occur slightly more often (1.14 times) in the PDT texts – see the column (7). On the other hand, the CLTT legal texts contain more constructions with ellipses (column (8)); they are about 1.8 times more frequent in CLTT than in PDT, which in fact goes against the explicitness requirement assumed in legal texts.

Finally, the statistics for both parenthetical constructions (column (9)), which appear about 2.6 times more often in legal texts, and constructions with numbers (column (10)), which occur about 4.3 times more often in CLTT, confirm the expected complex structure of legal texts.

The statistical comparison of administrative-legal texts in CLTT and the journalistic texts in PDT captured in Table 2 with regard to selected language phenomena relevant for the description of style and genre indicates and mostly confirms the expected complexity of legal text's sentence structure as reflected in its syntactic annotation. Therefore we believe our RExtractor system will be a huge help for the annotation of this kind of complex data.

GENRE	DESCRIPTION
advice	advice column, interpretation, instructions
caption	descriptions of pictures, graphs, tables
collection	collection of various texts in one document
comment	commentary on an actual topic (short)
description	description of a product, company, services
essay	larger report or comment (longer)
invitation	to concerts, exhibitions, etc.
letter	letters (from readers)
news	current news report
other	genre is uncertain - especially in isolated sentences
overview	list of currency rates etc.
interview	interview with a person, multiple topics
plot	description of a plot (film, TV program)
program	(cultural) program of TV, radio, exhibitions
review	critical review (books, films, exhibitions, concerts, theatre)
sport	sports news, results
survey	survey and its results
topic	topical interview, "actual conversation"
weather	weather forecast

Table 3: Genre categories annotated in the PDT 3.0 corpus, from (Mikulová et al., 2013).

4. Annotation of CLTT

With respect to the complexity of legal text sentences, we formulated an annotation scenario to make the process of arriving at a manually checked and corrected annotation of CLTT as simple and painless as possible, by using the following steps:

1. Tokenization and sentence segmentation
2. Complex sentence segmentation
3. Re-tokenization
4. Parsing CLTT using an automatic dependency parser
5. Manual correction of the parser output

We decided to apply this strategy for the following two reasons:

- accuracy of automatic parsers is better on shorter sentences
- annotators would check less nodes which makes annotation more comfortable and less erroneous.

4.1. Tokenization and Sentence Segmentation

Tokenization is the process of separating a text into meaningful units (tokens). Sentence segmentation is the process of separating a text into sentences, i.e. identifying sentence boundaries. We processed the CLTT texts by the standard tokenization and sentence segmentation procedures implemented in the Treex framework (Popel and Žabokrtský, 2010).

4.2. Complex Sentence Segmentation

We proposed an automatic procedure which splits long sentences. A *long sentence* is a sentence containing at least two segments. A *segment* is a part of a sentence between two numbering markers. It might not be a complete sentence nor even a complete clause. However, its manual annotation becomes more annotator friendly.

[8] Disregarding broadcast programs and weather.

CORPUS	LANGUAGE MEANS									
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
CLTT 1.0	27.66%	11.08%	0.71%	6.29%	41.67%	0.71%	6.29%	44.68%	20.92%	78.19%
PDT 3.0	5.03%	17.30%	0.25%	1.75%	9.56%	0.01%	7.16%	25.61%	8.09%	18.08%
GENRE										
advice	6.62%	19.13%	0.07%	0.60%	5.23%	0.07%	7.08%	23.36%	6.15%	15.02%
caption	2.76%	4.73%	0.39%	2.76%	9.86%	0.00%	7.89%	56.02%	11.64%	11.83%
collection	2.89%	12.96%	0.46%	2.52%	10.72%	0.00%	7.05%	32.89%	8.06%	27.58%
comment	6.18%	20.76%	0.16%	1.75%	10.05%	0.00%	7.43%	22.67%	7.21%	10.65%
description	6.06%	14.44%	0.25%	1.28%	10.85%	0.00%	7.66%	22.60%	6.98%	22.48%
essay	5.73%	17.47%	0.16%	1.02%	7.60%	0.01%	7.80%	23.76%	7.33%	9.97%
interview	4.35%	25.36%	0.20%	0.68%	4.15%	0.00%	5.44%	19.92%	4.01%	4.35%
invitation	4.76%	12.56%	0.37%	3.29%	12.32%	0.00%	10.37%	28.78%	12.80%	23.05%
letter	6.45%	23.04%	0.00%	3.00%	7.60%	0.00%	7.14%	30.18%	13.36%	16.13%
news	4.88%	19.68%	0.40%	2.74%	13.61%	0.01%	5.34%	20.77%	7.56%	18.74%
other	4.81%	14.09%	0.00%	0.86%	5.07%	0.00%	7.82%	32.22%	6.70%	16.32%
overview	3.32%	9.35%	0.30%	1.51%	5.43%	0.00%	14.18%	42.68%	11.01%	51.28%
plot	3.00%	6.00%	1.00%	4.00%	8.00%	0.00%	10.00%	45.00%	19.00%	22.00%
program	1.05%	3.35%	0.00%	1.05%	3.35%	0.00%	10.06%	81.55%	11.53%	82.39%
review	3.30%	9.73%	0.26%	2.19%	9.91%	0.00%	12.39%	33.53%	13.04%	11.19%
sport	4.00%	16.40%	0.16%	1.07%	6.32%	0.04%	6.81%	31.00%	11.61%	24.95%
survey	5.48%	15.93%	0.26%	1.83%	7.31%	0.26%	13.05%	38.38%	10.18%	30.55%
topic	6.75%	23.54%	0.08%	0.88%	5.06%	0.00%	5.67%	16.72%	4.29%	5.44%
weather	0.88%	3.54%	0.00%	0.00%	0.00%	0.00%	13.27%	80.53%	1.77%	69.03%

Table 2: The corpora CLTT 1.0 and PDT 3.0 and the following language means: (1) reflexive passive, (2) periphrastic passive, (3) chains of four genitive expressions, (4) chains of three genitive expressions, (5) chains of two genitive expressions, (6) construction with deverbative noun ending on *–ní*, *-tí* with genitive, (7) apposition, (8) ellipsis, (9) parenthesis, (10) numbers. The figures represent the proportion of a-nodes of a given language mean in a particular corpus. For example, 41.67% of the CLTT a-nodes (38,085 in total) are the heads of genitive phrases, like *shromažďování záznamů* (lit. *gathering records*). The proportion of such cases in PDT 3.0 (containing 1,957,247 a-nodes in total) is significantly lower, 9.56%. Out of all genres present in PDT 3.0, the news contain the highest number of genitive phrases, 13.61%.

ORIG	LONG SENTENCE	COMPL
s_1	(1) Complex sentence:	s_1n_1
	a) first subsection,	s_1n_2
	b) second subsection,	$s_1n_3m_1$
	1. paragraph,	$s_1n_3m_2$
	2. paragraph,	$s_1n_3m_3$
	c) third subsection.	s_1n_4
s_2	(2) Simple sentence.	s_2

Table 4: Original vs. complex sentence segmentation.

Table 4 illustrates the difference between the original (ORIG) and complex (COMPL) sentence segmentation.

Out of 1,128 sentences in CLTT, 101 sentences were identified as long sentences and we segmented them into 536 segments. The average sentence length of the non-segmented sentences (i.e., 1,027 sentences) is 25 tokens, while the long (i.e., segmented) sentences contain 91 tokens in average (17 sentences per segment). The longest sentence containing 491 tokens was split into 24 segments, the longest one contains 142 tokens.

4.3. Re-tokenization

We designed re-tokenization as a process of merging tokens. The standard tokenization splits all numbering types, e.g., it splits the string *(a)* into three tokens *(* and *a* and *)* that make the annotation more confused. We proposed a rule-based procedure for merging originally split numbering tokens back to one token. For illustration, see the node with the form *(7)* in Figure 1.

We handled references that refer either to other parts of the document or to a different document in the same way as numbering types. For illustration, see the node with the form *§18 odst. 3 zákona* in the tree displayed in Figure 1 for Example 1.

4.4. Parsing CLTT and Manual Correction

Both segments and non-segmented sentences were processed by the automatic parser (McDonald et al., 2005). Subsequently, the annotator processed the parser output:

- she checked both the tree structure and the analytical function assignment;
- she added inter-segment links for the nodes having their heads in a different segment – see the dotted arrows in Figure 2. This figure displays the dependency tree of the sentence segmented into four segments. According to the annotation guidelines, this sentence should be annotated as coordination of three predicates (i.e., *executes*, *transfers*, *fulfills*) where the comma in the segment s_1n_3 is its head.

5. Publishing CLTT

There are various ways of accessing the Czech Legal Text Treebank 1.0 which we have created in the course of the work described herein.

First, it can be downloaded from the LINDAT/CLARIN repository:

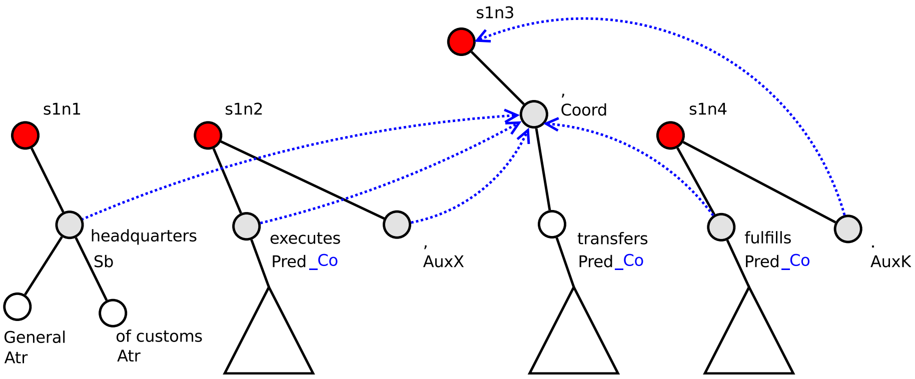


Figure 2: Illustration of merging segment annotations

http://hdl.handle.net/11234/1-1516

In addition, there are various tools for browsing and querying the treebank either locally or on-line - the TrEd graphical editor, the KonText KWIC search tool and PML TreeQuery, as described below.

5.1. TrEd editor

The users can view the (downloaded) treebank in the TrEd editor[9] that we used during the manual annotation. It is a fully customized and programmable graphical editor and viewer for tree-like structures. Among other projects, it was used as the main annotation tool for annotations in the Prague Dependency Treebank.

We implemented a new TrEd extension called *INTLIB Annotation*. This extension can be installed directly in TrEd using *Setup* → *Manage Extensions* → *Get New Extensions*. It offers several new features:

- running new macros for more comfortable annotation
- tracking changes in a tree structure and in a-node attributes made by annotators
- making inter-segment links

5.2. KonText

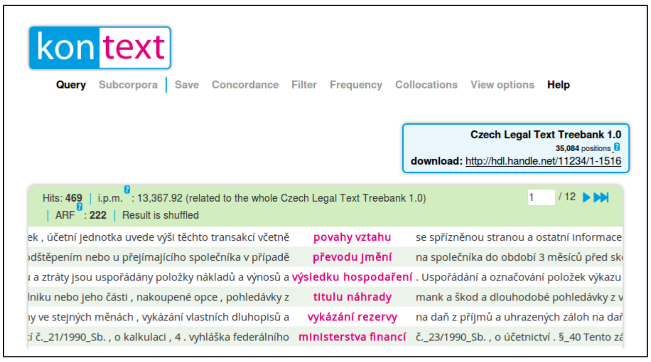
KonText[10] is a web application for querying corpora on-line within the LINDAT/CLARIN project. The users can evaluate simple and complex queries, display their results as concordance lines, compute frequency distribution, calculate association measures for collocations and do further work with the data.

5.3. Tree Query

Tree Query[11] is a powerful open-source search tool for all kinds of linguistically annotated treebanks available on-line within the LINDAT/CLARIN project. The users can evaluate complex tree queries and display their results graphically highlighted in the dependency trees. Tree Query can also be used from within the TrEd editor.

6. Future Work

We consider to include the Czech Legal Text Treebank within the Universal Dependencies framework (Nivre et al., 2016), but we do not plan to enlarge it at this time. Instead, we will focus on experimenting with various parsers in order to improve both performance of legal text parsing and, most importantly, the RExtractor performance.



Czech Legal Text Treebank 1.0
35,084 positions
download: http://hdl.handle.net/11234/1-1516

Hits: 469 | i.p.m. : 13,367.92 (related to the whole Czech Legal Text Treebank 1.0) 1 / 12
| ARF : 222 | Result is shuffled

ek , účetní jednotka uvede výši těchto transakcí včetně	povahy vztahu	se spřízněnou stranou a ostatní informace
odštěpením nebo u přejímajícího společníka v případě	převodu jmění	na společníka do období 3 měsíců před sk
a ztráty jsou uspořádány položky nákladů a výnosů a	výsledku hospodaření	. Uspořádání a označování položek výkazu
lniku nebo jeho části , nakoupené opce , pohledávky z	titulu náhrady	mank a škod a dlouhodobé pohledávky z v
ny ve stejných měnách , vykázání vlastních dluhopisů a	vykázání rezervy	na daň z příjmů a uhrazených záloh na daň
í č._21/1990_Sb. , o kalkulaci , 4 . vyhláška federálního	ministerstva financí	č._23/1990_Sb. , o účetnictví . §_40 Tento zá

Figure 3: The occurrences of genitive expressions in CLTT presented as concordances in the KonText on-line service.

7. Conclusions

We introduced a new member of the family of Prague dependency treebanks. The Czech Legal Text Treebank 1.0 is a morphologically and syntactically annotated corpus of 1,128 sentences. The treebank contains texts from the legal domain, namely the documents from the Collection of Laws of the Czech Republic. The treebank presents a unique and interesting language resource.

Legal texts differ from other domains in several language phenomena. We compared the treebank with the largest annotated corpus available for Czech, namely the Prague Dependency Treebank containing mostly newspaper texts. Sentences in legal texts are typically long and very complex and it makes both their manual annotation and parsing more difficult. We have described our strategy for handling long sentences by segmenting them and recombining them back after parsing.

8. Acknowledgments

We gratefully acknowledge support from the Technology Agency of the Czech Republic (grant no. TA02010182) and from SVV project no. 260 224 and from the LINDAT/CLARIN Research Infrastructure project of the Ministry of Education, Youth and Sports of the Czech Republic No. LM2015071. This work has also been using language resources developed and distributed by the LINDAT/CLARIN project.

[9] http://ufal.mff.cuni.cz/tred/

[10] https://lindat.mff.cuni.cz/services/kontext

[11] https://lindat.mff.cuni.cz/services/pmltq

9. Bibliographical References

John Gibbons, 2008. *Language and the Law*, pages 285–303. Blackwell Publishing Ltd.

Bohuslav Havránek. 1932. *Spisovná čeština a jazyková kultura*. Praha: Melantrich.

Milan Jelínek. 1995. Kultura jazyka a odborný styl. In *Termina 94*, pages 7–29, Liberec: PFTU.

Milan Jelínek. 1996. Styl administrativně-právní. In *Jazyk a jeho užívání. Sborník k životnímu jubileu prof. O. Uličného*, pages 240–250, Praha.

Jan Kořenský. 1989. Právní jazyk, právní komunikace a interpretace. In *Stát a právo 27*, pages 33–39, Praha.

Vincent Kríž and Barbora Hladká. 2015. RExtractor: a robust information extractor. In Matt Gerber, Catherine Havasi, and Finley Lacatusu, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, Denver, CO, USA. Association for Computational Linguistics.

Vincent Kríž, Barbora Hladká, Martin Nečaský, and Tomáš Knap. 2014. Data extraction using NLP techniques and its transformation to linked data. In *Human-Inspired Computing and Its Applications - 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I*, pages 113–124.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, BC, Canada. Association for Computational Linguistics, Association for Computational Linguistics.

Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, and Zdeněk Žabokrtský. 2013. From PDT 2.0 to PDT 3.0 (modifications and complements). Technical Report ÚFAL TR-2013-54.

Eva Minářová, Marie Krčmová, Jan Chloupek, and Marie Čechová. 2003. *Současná česká stylistika*. ISV nakladatelství, Praha, 1 edition.

Martin Nečaský, Tomáš Knap, Jakub Klímek, Irena Holubová, and Barbora Hladká. 2013. Linked open data for legislative domain - ontology and experimental data. In *Lecture Notes in Business Information Processing*, pages 172–183, Berlin / Heidelberg. Springer.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

Petr Pajas and Jan Štěpánek. 2006. XML-based representation of multi-layered annotation in the PDT 2.0. In Richard Erhard Hinrichs, Nancy Ide, Martha Palmer, and James Pustejovsky, editors, *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47, Genova, Italy. ELRA, ELRA.

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. 2013. Introducing the prague discourse treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan. Asian Federation of Natural Language Processing, Asian Federation of Natural Language Processing.

Lucie Poláková, Pavlína Jínová, and Jiří Mírovský. 2014. Genres in the prague discourse treebank. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1320–1326, Reykjavík, Iceland. European Language Resources Association.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Peter Meijes Tiersma. 1999. *Legal language*. University of Chicago Press.

10. Language Resource References

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague dependency treebank 3.0. `http://ufal.mff.cuni.cz/pdt3.0`.

Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). The syn-series corpora of written czech. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).