Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

HABILITATION THESIS

Barbora Vidová Hladká

Creating and Exploiting Annotated Corpora

Computer Science
Mathematical Linguistics

Prague 2019

Contents

1. Research Overview

This thesis focuses on creating and exploiting annotated linguistic corpora in the field of computational linguistics, including their use in applications and in computer-assisted learning. While annotated corpora are in the center of the research on linguistic resources in general, this thesis focuses on various aspects of their morphological and syntactic annotation.

From the historical perspective, the first two morphologically and in part syntactically annotated corpora appeared in the late 1960s, and both are related to Czech linguists. The `Brown Corpus` for English (Francis and Kucera, 1964) has been conceived by a Czech emigrant, the linguist Henry Kučera, and the `Czech Academic Corpus` has been created in the Institute of the Czech Language in Prague by the team of Marie Těšitelová (Králík and Uhlířová, 2007). Already in my diploma thesis and then in my dissertation, I have used the latter as the training data for the first statistical tagger(s) of Czech. This unique position of Czech and English has been later strengthened by two other annotation projects, namely the `Prague Dependency Treebank` and the `Penn Treebank` (Marcus et al., 1993). I highly appreciate the opportunity to having taken part in the `Prague Dependency Treebank` project since its very beginning, mainly being the coordinator of its morphological annotation and the co-author of the most cited publication on this corpus (Hajič et al., 2003).

Both these annotation projects and experiments carried out on the data they resulted in inspired me to focus on several research questions over the following three research areas: Academic Corpus Annotation, Alternative Corpus Annotation, and Information Extraction. I am submitting this thesis, "Creating and Exploiting Annotated Corpora", to present results that I have achieved in these three areas since I completed my Ph.D. degree. In Chapter 2 and in the relevant papers in Appendix A, I describe my contribution to the Academic Corpus Annotation projects which I coordinated. Such projects run according to a three step scenario: (i) selection of texts to be annotated, (ii) formulation of annotation guidelines by linguists, and (iii) annotation of the texts by trained annotators. Typically, they are extremely expensive in terms of human resources, annotation duration, and their budget. It motivated me to explore a different strategy, namely to acquire such annotation as a by-product of so-called Alternative Annotation having the form of on-line games and school sentence diagramming , which I elaborate in Chapter 3 and in the relevant papers in Appendix B. While annotated corpora are of great importance to linguistic research, they are also indispensable for the development of text analysis tools and applications by using machine learning methods trained on the corpora; in Chapter 4 and Appendix C, I focus on the task of Information Extraction in the legal and environmental domains.

Each of the aforementioned chapters is introduced with the relevant research questions together with a summary of its topic and is closed with a summary of my contribution in the given areas.

In the rest of the present Chapter, I summarize my research in the three aforementioned areas, and provide a list of additional results and publications that I authored or co-authored (often with my students) and which complement the main ones this thesis consists of (listed in bold).

1.1 Terminology

This thesis is concerned with the creation and use of annotated corpora. For easier understanding of its content, it is useful to begin with a brief description of the main concepts.

A *corpus* is a collection of language data compiled from either written texts or transcriptions of recorded speech. *Corpus annotation* is the process of adding linguistic information to a corpus that afterwards becomes an *annotated corpus*. The purpose of creating both corpora and annotated corpora is to create an objective evidence of the real usage of the language.

Corpus annotation can be undertaken at different levels of linguistic analysis. For illustration, (1) the most common type of annotation is labeling words by tags indicating their part-of-speech classes; (2) at the morphological level words are annotated for their morphological features in addition to their part-of-speech classes, e.g., case, gender, number of nouns. The morphological annotation is known as *tagging*; (3) at the syntactic level sentences are annotated for their syntactic structure. The syntactic annotation is known as *parsing*; (4) at the discourse level words can be annotated to show coreference links in a text, e.g., the link between the pronoun *he* and the proper name *Forman* in: *Douglas convinced Forman to show Nicholson something, which he did.*

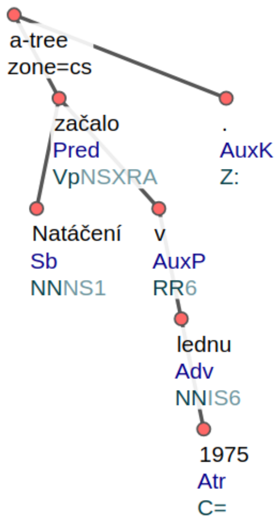

Figure 1: Illustration of the `Prague Dependency Treebank` annotation scheme on the sentence *Natáčení začalo v lednu 1975.* (*Filming began in January 1975.*)

A linguistic analysis is encoded in an *annotation scheme*. Figure 1 illustrates the annotation scheme of the `Prague Dependency Treebank`, the most distinguished Czech annotated corpus: (i) the syntactic analysis is encoded using a *dependency tree* where each arc represents a grammatical dependency categorized according to the role of the dependent word, e.g., the verb with the role of predicate is understood as the centre of the sentence, see the node *začalo* with the analytical function `Pred`(icate). As syntactically annotated corpora often appear similar to tree structures, they are known as *treebanks*. (ii) the morphological analysis is encoded using a *tag* that is a string of characters encoding morphological categories, e.g., see the node *lednu* and its morphological tag `NNIS6` for the locative case of a singular masculine inanimate noun.

Corpus annotation can be achieved manually by annotators, fully automatically by procedures, or semi-automatically as an interaction between annotators and procedures. *Gold standard corpora* are manually annotated corpora of high quality and they are essential for training and evaluation of statistical machine learning algorithms, *corpus-based procedures*. The standard methodology of evaluation is to apply procedures to a test set taken from a golden corpus and compare their annotation to the gold annotation in the corpus. For example, a dependency parsing procedure generating dependency trees can be evaluated with the *F1 score* defined as the harmonic mean of precision and recall, where *precision* is the the number of correct arcs out of the number of arcs in the generated tree and *recall* is the number of correct arcs out of the number of arcs in the gold corpus.

Tagging and parsing are examples of problems in *Natural Language Processing* (NLP) that is concerned with "the design and implementation of effective natural language input and output components for computational systems" according to (Dale et al., 2000).

1.2 Academic Corpus Annotation

To create a gold standard corpus, multiple experts annotate the texts independently and the inter-annotator agreement is computed to ensure quality of annotations. I call this process an *academic corpus annotation* and in the next I deal with the following gold standard corpora: `Czech Academic Corpus` (`CAC`), `Czech Legal Text Treebank` (`CLTT`), `Prague Dependency Trebank` (`PDT`), `STYX`.

Czech Academic Corpus

The process and length of the `CAC` projects conflicts with every embedded notion of a traditional project. Its morphological and syntactic annotation in 1971-1985 was a pioneering effort, like the annotation of the `Brown corpus` of American English (Francis and Kucera, 1964), the `LOB corpus` of British English (Atwell et al., 1984), and the `Talbanken corpus` of Swedish (Einarsson, 1976a), (Einarsson, 1976b). Fortunately the electronic version of `CAC` was successful in keeping up with the fast technical progress, mainly changes of data media. Hence in the early 1990s when statistical approaches in NLP have become dominant I could use `CAC` for the very first statistical learning experiments on the tagging of Czech texts (Hladká, 1994). Ten years later I was the principal investigator of the project "Resources and Tools for Information Systems" the main goal of which was a transformation of the original `CAC` annotation scheme into the annotation scheme of `PDT` dominating the annotation projects on Czech since 1990s.[1] I greatly appreciate having Jan Králík in the team since he was a member of the original team. Given that we together could finally document the first stage of `CAC` (Hladká and Králík, 2006). After the initial transformation steps, `CAC version 1.0` was published as a monograph accompanied by the medium with the data and tools (Hladká et al., 2007). The data and tools of `CAC 2.0` were published by the Linguistic Data Consortium (Hladká et al., 2007) separately from

[1]2004-2008, 1ET101120413, funded by the Grant Agency of the Czech Academy of Sciences.

their guide (Hladká et al., 2008a). Alla Bémová and Zdeňka Urešová participated in the transformation of CAC as well and we studied syntactic properties of its spoken texts in a follow-up research **(Hladká et al., 2011)**. Recently, CAC has reached another significant milestone when Dan Zeman and I transformed the CAC 2.0 annotations into the Universal Dependencies (UD) annotation scheme (Nivre et al., 2017).

Corpus-Based Exercise Book of Czech

The years of intensive work with CAC and PDT inspired me to explore an innovative idea of using annotated corpora in language classes where morphology and syntax are taught and practiced using sentence diagrams. Ondřej Kučera, a Master student of mine, implemented the Styx system in his Master thesis (Kučera, 2005). It is an exercise book of 11 thousand Czech sentences that can be used to practice sentence diagramming with key answers available. The sentences in Styx were selected from PDT and their annotations were transformed into sentence diagrams taught at Czech language classes. We demonstrated Styx internationally for the first time at the prestigious conference HLT/EMNLP in 2005 and we took third place in the interactive demonstration session (Hladká and Kučera, 2005). Later on, I was a co-PI of Ondřej's project "Prague Dependency Treebank as an exercise book of Czech" and we documented Styx in great details at the end of this project **(Kríž and Hladká, 2008)**.[2]

The sentence diagrams have been available only while running Styx and thus any linguistic and practical studies were impossible. Therefore Karolína Kuchyňová, a student of mine, applied the transformation rules outside the Styx system and we published the annotated corpus STYX 1.0 where both the original PDT annotations and the sentence diagrams are available (Hladká et al., 2017).

Czech Legal Text Treebank

CLTT is a morphologically and syntactically annotated corpus of legal texts. Its volume is many times smaller than the volumes of PDT and CAC and also our motivation to create another Czech treebank was different.

A parsing procedure is a substantial component of the RExtractor system that I have been developing with Vincent Kríž, a PhD student of mine. The legal domain presents a target domain of this system and since no in-domain legal gold standard treebank exists to train a parsing procedure (*parser*) on we apply a cross-domain approach to parse legal texts using the procedure trained on newspapers. We created CLTT to evaluate the parsing procedure of Czech employed by RExtractor.

At least to my knowledge, no other morphologically and syntactically annotated corpus in the legal domain exists and here is my explanation: one has to be very patient and highly organized to manipulate dependency trees that are bigger than the screen. This happens for sentences in legal documents often. Such an annotator is only one and her name is Zdeňka Urešová.

In total, we published two versions of CLTT: CLTT 1.0 (Kríž et al., 2015), (Kríž et al., 2016) contains the morphological and syntactic annotations that we en-

[2]2007-2008, GAUK 55907/2007, funded by the Grant Agency of Charles University.

riched with the annotation of entities and relations in `CLTT 2.0` (Kríž and Hladká, 2017), **(Kríž and Hladká, 2018)**. Both versions were built in the project "Intelligent Library" (INTLIB) where I coordinated the team for natural language processing.[3] Not only `CAC 2.0` but also `CLTT 2.0` undertook the transformation into the `UD` annotation scheme which leads me to highlight the fact that among the languages having corpora annotated using this annotation scheme, Czech is the winner with 2,222K annotated tokens.

All of the mentioned corpora contain texts produced by native speakers of Czech. Texts written by non-native speakers pose a challenge for natural language processing as well and I am happy to recall the uniqueness of Czech, this time thanks to `CzeSL`, a corpus of learner Czech. Jirka Hana is one of its authors and we together have started its manual annotation using the `UD` annotation scheme. We revised the annotation guidelines formulated for texts of native speakers to annotate texts of non-native speakers (Hladká and Hana, 2017), (Hana and Hladká, 2018) and published the annotated portion of `CzeSL` (Hana and Hladká, 2019).

1.3 Alternative Corpus Annotation

The academic corpus annotation is expensive in terms of time and funding. The annotation experience I have acquired while participating in the academic annotation projects provided me with a strong evidence to care whether we can get annotated corpora in a less demanding process. I was interested in alternative ways of annotation gathered mostly in the model of *crowdsourcing* that engages typically a large number of individuals to achieve a given goal. No doubt the Internet and schools are the most appropriate environments to implement crowdsourcing projects.

Luis von Ahn, a pioneer in the field of human computation, invented on-line *Games With A Purpose* (GWAPs) that can be briefly and easily characterized as follows: (1) GWAPs are designed to annotate data for those tasks that have not achieved human performance yet; (2) Users play GWAPs and produce annotated data as a by-product of enjoyment (von Ahn, 2006), (von Ahn and Dabbish, 2008). The very first game that his team implemented was the ESP game where an image is shown to two players and they label it with words they expect the opponent will use (von Ahn, 2006).[4] This game have met with a great success when the players labeled incredible amount of images in a short time, as evidenced by e.g., "*as of July 2008, 200,000 players had contributed more than 50 million labels*" in (von Ahn and Dabbish, 2008). No wonder, then, that such success inspired other researchers, including me, to implement GWAPs for various fields of study, including Natural Language Processing. Undoubtedly texts are more complicated than images in terms of a period of time taken to understand their content. That is naturally longer for texts, so it is necessary to think carefully how to submit texts into game sessions to attract as many players as possible.

Since the manual coreference resolution shows a substantially higher performance than the automatic procedures (Sukthanker et al., 2018), we implemented

[3] 2012-2015, TA02010182, funded in 2012-2015 by the Technology Agency of the Czech Republic.

[4] `http://en.wikipedia.org/wiki/ESP_game`

the `PlayCoref` game where players connect all co-referring words in as many sentences as possible, e.g., *Forman* and *he* and *Nicholson* and *actor's* in: *Douglas convinced Forman*$_1$ *to show Nicholson*$_2$ *something, which he*$_1$ *did and restored the actor's*$_2$ *confidence*, see **(Hladká et al., 2009a)**. A workflow of the game was designed to be language independent and we demonstrated it on Czech and English. Playing `PlayCoref` requires only a basic knowledge of the language of the game; no extra linguistic knowledge is required **(Hladká et al., 2009b)**. Given that, more attention has to be paid to checking the annotations coming from the game sessions (Hladká et al., 2011). Independently of the `PlayCoref` development, Chamberlain et al. (2008) implemented `Phrase Detectives`, a GWAP on anaphoric links that are closely interrelated with coreference links collected by `PlayCoref`. Both games are the pioneers among the textual GWAPs.

In addition to `PlayCoref`, I explored an innovative way to engage crowds of students to enlarge the volume of a morphologically and syntactically annotated corpus as follows: students create sentence diagrams electronically for the sentences of their choice and send them to an academic team who transforms them into a target annotation scheme. Jirka Hana, I, and Ivana Lukšová, a PhD student of mine, implemented the `Čapek` editor to enable drawing of sentence diagrams electronically (Hana and Hladká, 2012). The transformation of Czech sentence diagrams into the `PDT` annotation scheme is reverse to the one implemented in the `Styx` system. Since it is not deterministic we included the combination of sentence diagrams into the transformation rules **(Hana et al., 2014)**. Marie Konárová, one of my Master students, organized collecting of sentence diagrams created by teachers and students using `Čapek` (Konárová, 2012) and Karolína Kuchyňová analyzed the diagrams that Marie collected (Kuchyňová, 2016).

1.4 Information Extraction

Most NLP procedures use machine learning methods, typically statistical- or neural network-based, in order to analyze (process) new texts. The basic assumption of these methods is that a relevant gold standard corpus exists. I participated in the above mentioned INTLIB project on a complex NLP procedure, namely on the task of information extraction.

The INTLIB project of applied research was addressed by Sysnet Ltd. and two departments of the Faculty of Mathematics and Physics, namely the Department of Software Engineering (KSI) and the Institute of Formal and Applied Linguistics (ÚFAL). Its aim was to provide a more efficient and user-friendly tool for querying textual documents than full-text search. On the input a collection of documents from the legal and environmental domains was assumed, namely acts in the accounting subdomain and environmental impact assessment reports, resp. In the first phase, under my supervision, a knowledge base was extracted from the documents using natural language processing tools; in the second phase, under the supervision of Martin Nečaský from KSI, the extracted knowledge was represented in the framework of Linked Open Data, see (Nečaský et al., 2013), (Holubová et al., 2014), (Kríž et al., 2014). My NLP group approached the task of querying textual documents as the task of information extraction.

Information Extraction (IE) extracts structured data from unstructured text by identifying entities and relations between them. Research in IE has been shaped by the series of competition-based Message Understanding Conferences in 1987-1998 (Grishman and Sundheim, 1996), (Chinchor, 1998) and a variety of approaches to constructing IE systems has been developed since that time. We explored and refined the one that pre-process input texts with NLP tools (Mooney and Bunescu, 2005). We, I and Vincent Kríž, have been developing `RExtractor` that is a system of information extraction for knowledge base construction in the legal domain. A knowledge base consists of entities related to the accounting domain and the relations of definition, right, and obligation. `RExtractor` processes the input documents by the procedures of tagging and parsing and consequently queries the dependency trees of the input sentences to extract the knowledge base **(Kríž and Hladká, 2015)**. The queries are based on the entity and relation annotation in `CLTT 2.0`. For illustration, `RExtractor` extracts the underlined entities in the sentence *Accounting units shall take inventory of their assets and liabilities pursuant to section 29 and 30.* In addition, the relation of obligation is extracted: what *accounting units* has to do – *take inventory.* We also studied the influence of the parsing procedure on the `RExtractor` performance. In particular, the high frequency of long and complex sentences cause problems to the parsing procedure and therefore we performed a study, in which we split sentences into smaller segments and verified whether their automatic parsing and subsequent joining of their dependency trees would contribute to an improvement of the dependency trees of the original sentences **(Kríž and Hladká, 2016)**. Bohdan Maslowski, a Master student of mine, processed documents from the legal domain as well (Maslowski, 2015). Bohdan experimented with court decisions and his aim was to identify persons and their roles in a given case using both machine learning and rule-based methods.

The Environmental impact assessment (EIA) is a formal process used to predict the environmental consequences of a construction plan. The genre of EIA documents differs from the documents that we mined in the legal domain. We, I and Ivana Lukšová, implemented the `EIA extractor` system for extraction of quantitative data from the Czech EIA documents (Lukšová and Hladká, 2015). This is a rule-based system encoding regular expressions on morphological information, e.g., we use the pattern `(Numeral) (Adjective Genitive)? (Noun Genitive)` to get the number of parking slots from the sentence *Bude tam 1 150 parkovacích míst (There will be 1,150 parking slots).* Our extraction and representation methodology has been certified by the Ministry of the Environment of the Czech Republic (Borš et al., 2014).

Both the legal and the environmental domains are still largely underrepresented in the NLP literature despite its potential for generating interesting research questions. This fact and a multi-disciplinary nature of the INTLIB project make our work original.

Selected publications

Academic corpus annotation

1. Hladká B., Bémová A., Urešová Z. Syntaktická proměna Českého akademického korpusu. *Slovo a slovesnost*, No. 4, Ústav pro jazyk český AV ČR, ISSN 0037-7031, pp. 268-287, 2011. [**my share 40 %**]
   DOI: –

2. Hladká B., Kučera O. An Annotated Corpus Outside Its Original Context: A Corpus-Based Exercise Book. In: *Proceedings of the ACL-08: HLT Third Workshop on Innovative Use of NLP for Building Educational Applications*, The Ohio State University, Columbus, Ohio, USA, pp. 36-43, 2008. [**70 %**]
   DOI: `https://doi.org/10.3115/1631836.1631841`

3. Kríž V., Hladká B.: Czech Legal Text Treebank 2.0 In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association, Paris, France, ISBN 979-10-95546-00-9, pp. 4501-4505, 2018. [**65 %**]
   DOI: –

Alternative annotation

4. Hladká B., Mírovský J., Schlesinger P. Designing a Language Game for Collecting Coreference Annotation. In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, Association for Computational Linguistics, Suntec, Singapore, ISBN 978-1-932432-52-7, pp. 52-55, 2009. [**70 %**]
   DOI: `https://doi.org/10.3115/1698381.1698389`

5. Hladká B., Mírovský J., Schlesinger P. Play the Language: Play Coreference. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Association for Computational Linguistics, Suntec, Singapore, ISBN 978-1-932432-61-9, pp. 209-212, 2009. [**70 %**]
   DOI: `https://doi.org/10.3115/1667583.1667648`

6. Hana J., Hladká B., Lukšová I. Sentence diagrams: their evaluation and combination. In: *Proceedings of the 8th Linguistic Annotation Workshop*, Association for Computational Linguistics, Dublin City University, Dublin, Ireland, ISBN 978-1-941643-29-7, pp. 38-47, 2014. [**60 %**]
   DOI: `https://doi.org/10.3115/v1/W14-4905`

Information extraction

7. Kríž V., Hladká B. RExtractor: a Robust Information Extractor. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics, Denver, CO, USA, ISBN 978-1-941643-49-5, pp. 21-25, 2015. [**70 %**]
   DOI: `https://doi.org/10.3115/v1/N15-3005`

8. Kríž V., Hladká B. Improving Dependency Parsing Using Sentence Clause Charts. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop*, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-945626-02-9, pp. 86-92, 2016. [**60 %**]
   DOI: `https://doi.org/10.18653/v1/P16-3013`

Acknowledgement

I would like to express my deep gratitude to all of my closest collaborators, both researchers and students. I have already introduced most of them either directly by their names or indirectly as the co-authors of the cited works. Unfortunately this list is incomplete and I wish to add a few colleagues.

I can not imagine a better way to start my research after finishing my PhD study than to spend one year at the Center for Language and Speech Processing of Johns Hopkins University in Baltimore, Maryland, USA. It could not happen without Fred Jelinek (†2010), a highly recognized Czech-American researcher and the director of the Center in 1993-2010.

I have the honour of doing research next to an outstanding personality of Eva Hajičová. She is a great source of inspiration what I illustrate using one example: Eva directed me and Jan Václ, a Master student of mine, to explore annotated corpora towards research focusing on text and discourse. We followed her notion of salience that studies dynamic of discourse and we experimented with machine learning of salience (Václ, 2015), (Zikánová et al., 2015).

Martin Holub is a colleague of mine not only in teaching but also in addressing the very interesting task of Native Language Identification where the goal is to predict an author's native language based only on his/her production in a second language. We participated in two shared tasks having English as a second language. The 2013 shared task had written inputs (Hladká et al., 2013) while the 2017 shared task combined both written and spoken inputs. We participated in the latter shared task with the group of Pavel Ircing from the Department of Cybernetics of the University of West Bohemia in Plzeň, Czech Republic and we took first place (Ircing et al., 2017).

2. Academic Corpus Annotation

In the present thesis we are concerned with annotated corpora that are collections of texts enriched with linguistic information useful for both theoretical and empirical linguistic research. We focus on the academic annotation of three Czech corpora, namely the `Czech Academic Corpus` (`CAC`), the `Czech Legal Text Treebank` (`CLTT`), and `STYX`. They are diverse in many ways, which has made their annotation and exploration exciting: while `CLTT` was annotated from scratch, the original annotations of `CAC` and `STYX` were transformed to other annotation schemes. The annotation scheme of the `Prague Dependency Treebank` (`PDT`), the highly appreciated annotated corpus of Czech (Hajič et al., 2018), served as the target scheme of `CAC` and `CLTT` and as the source scheme of `STYX`. Very recently, the three corpora undertook the transformation into the `Universal Dependencies` (`UD`) annotation scheme that currently dominates the annotation projects all over the world. The three corpora are available in the `LINDAT/CLARIN` repository[1] and are searchable on-line using the `KonText` corpus manager[2] and the `PML-TQ` search tool.[3] Table 1 provides the stories of `CAC`, `CLTT`, and `STYX` in a nutshell.

corpus	`CAC`	`CLTT`	`STYX`
source annotation scheme	original `CAC`	×	`PDT`
target annotation scheme	`PDT`, `UD`	`PDT`, `UD`	sentence diagrams, `UD`
interesting issues	missing tokens, spoken texts	complex tokens, complex sentences	complex nodes
exploitation of the corpus	tagging, parsing	cross-domain parsing, information extraction	tagging, parsing

Table 1: Creating and exploiting annotated corpora in a nutshell.

2.1 Czech Academic Corpus

> **Supervised machine learning methods need training data. What should be done with the corpus annotated thirty years ago in order to include it in the present corpus?**

The `Czech Academic Corpus` is an annotated corpus that significantly contributed to the beginning of Czech language processing in the early 1990s. We used it as training data in the very first experiments on Czech language tagging using statistical methods (Hajič and Hladká, 1997). Its annotation was conducted in the De-

[1]`http://lindat.mff.cuni.cz/repository`
[2]`http://lindat.mff.cuni.cz/services/kontext/`
[3]`http://lindat.mff.cuni.cz/services/pmltq`

partment of Mathematical Linguistics of the Czech Language Institute of the Academy of Sciences of the Czech Republic in 1971-1985. In total, 540,000 tokens of the written and spoken texts (transcriptions) were manually morphologically and syntactically annotated. From todays' perspective it is useful to mention that the corpus annotation was a secondary goal of the quantitative analysis of Czech (Těšitelová, 1985). This is underscored by the fact that the `CAC` annotation was not documented by its team. It was done many years later by Hladká and Králík (2006) and Králík and Uhlířová (2007).

From both historical and international perspectives, we consider `CAC` and three other corpora to be pioneering annotated corpora: (i) the `Brown corpus` of American English written texts of 1 million tokens annotated with their part-of-speech classes (Francis and Kucera, 1964), (ii) the `LOB corpus` (Atwell et al., 1984), a British version of the Brown corpus, and (iii) the `Talbanken` corpus of written and spoken Swedish texts of 350,000 tokens annotated with their part-of-speech classes and phrase structures (Einarsson, 1976a), (Einarsson, 1976b). At the end of the 20th century, many annotation projects started and the pioneering corpora faced a new situation in terms of their internal formats and annotation schemes. Their transformation seemed to be necessary: the `Talbanken` transformation covered the internal format conversion only (Nilsson et al., 2005); the part-of-speech tagset of the `Brown corpus` was modified in the project of `Penn Treebank` and the `Brown corpus` consequently became its part (Marcus et al., 1993).

The annotation project of `PDT` was a main motivation for getting back to `CAC` in order to increase the volume of Czech data needed for supervised machine learning procedures, mainly tagging and parsing. The Functional Generative Description views language as a system of levels (Sgall et al., 1986). It was used as the theoretical framework of the `PDT` annotation scheme which is firmly anchored in three basic annotation layers, technical counterparts of the levels from the underlying theory: (1) the morphemic m-layer with detailed part-of-speech tags and rich morphological information, (2) the syntactic a-layer having the form of dependency tree with the verb as the root of the tree and with relations labeled by analytical functions such as Subject (`Sb`), Object (`Obj`), Adverbial (`Adv`), etc., and (3) the underlying dependency-based syntactic t-layer with dependency tree structures labeled by functors such as Actor, Patient, Addressee, etc. The first two layers are illustrated on Example (1) in Figure 2 .

(1) *Sečíst pouhým okem stranickou příslušnost zvednutých rukou bylo ve dvousetčlenné Poslanecké sněmovně nemožné.*

Lit. *It was impossible to count the party's affiliation of raised hands in the two-hundred member Chamber of Deputies with the naked eye.*

The `CAC` transformation was carried out as a transformation of both the internal format and the annotation schemes according to `PDT 2.0`. It was almost a detective job mainly because of the brevity of the annotation scheme description and unavailability of the annotation instructions. The handwritten and spoken sources were not available as well and that made the transformation more intricate. `CAC 1.0` came to fruition as a result of the transformation of the internal format and morphological annotations that we organized using an automatic procedure (Hladká et al., 2007). The pilot study (Ribarov et al., 2006) showed that the manual syntactic annotation of the written texts is more effective than the annotation transformation. For that reason, `CAC 2.0` is a joined result of the checking of the morphological annotations in `CAC 1.0` and the manual syntactic annotation of the written texts (Hladká et al., 2008a), (Hladká et al., 2008b). Finally, we implemented an automatic procedure to transform `CAC 2.0` into `UD` (Nivre et al., 2017).

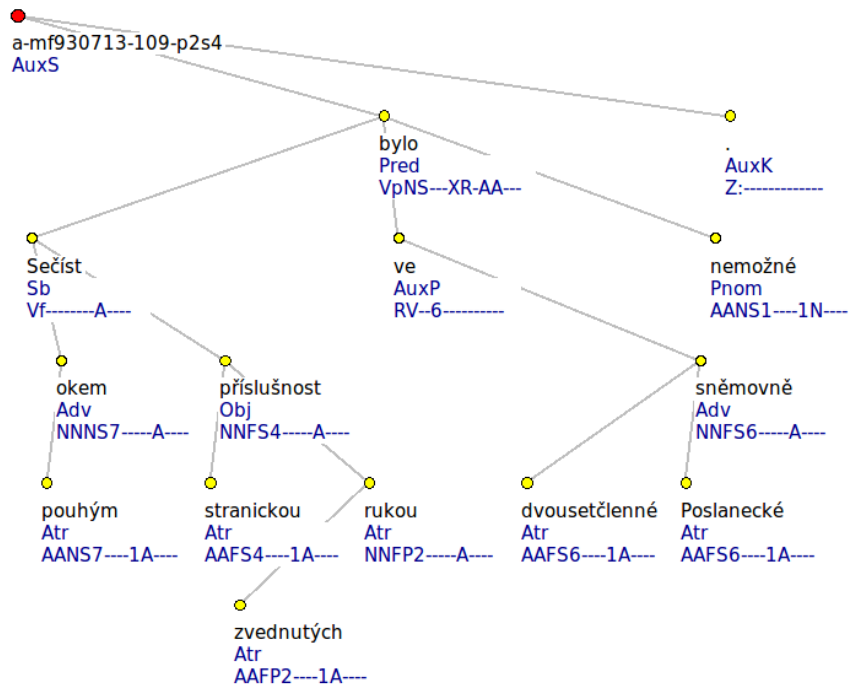

Figure 2: Example (1) annotated according to the `Prague Dependency Treebank` annotation scheme.

The most challenging issues we have dealt with concern different annotation instructions for particular language phenomena and the syntactic annotation of spoken texts. Although the annotation frameworks of `CAC` and `PDT` have grown out of the same European syntactic tradition (Šmilauer, 1947), there exist language phenomena that are handled either differently in both corpora or are handled in one of the corpora only, e.g., (i) some tokens were omitted from the `CAC` texts. These omissions included tokens that can be read differently (e.g., digit tokens *34* vs. *thirty-four* vs. *four-and-thirty*) or tokens that are not pronounced at all (punctuation). To be compatible with `PDT`, we manually inserted specific elements to show missing tokens in the texts; (ii) the syntactic layer in `CAC` has two sub-layers: intra-clausal that corresponds to the `PDT` a-layer and inter-clausal that captures syntactic relations between clauses in a complex sentence; (iii) the `PDT` a-layer fully reflects the surface structure of a sentence and places prepositions as roots of prepositional phrases. However, `CAC` does not include prepositions in the syntactic structure; (iv) conjunctions and reflexive particles are included in the syntactic structures in both corpora but according to different instructions.

Once we published `CAC 2.0` we started to think about a syntactic transformation of the spoken texts the structure of which is of a very specific character. The contemporary Czech syntactic theory is entirely based on written Czech, from which the spoken form is substantially different (Müllerová, 1994). We were wondering whether it is possible (i) to apply the experience and knowledge gained from the annotation of the written texts to the spoken texts; (ii) to apply the `PDT` annotation instructions directly to the `CAC` spoken texts or if *some* modification of the instructions is needed; (iii) to annotate the spoken texts as they are or if *some* editing of their sentence structure must precede the annotation itself. We analyzed the `CAC` tags designed for spoken texts and we found out that their repertoire was rather limited. There are tags for (1) non-identical word repetition, (2) identical word repetition, and (3) ellipsis illustrated in Example (2) where the lower indices indicate these three phenomena.

However, a detailed analysis of the spoken part of CAC showed that phenomena very well known from spoken speech (Fitzgerald, 2009) are present there but were not annotated, e.g., filler words, incoherent utterances. The PDT annotation framework does not cover spoken language structures, mainly because the PDT does not contain spoken texts. Given the above-mentioned arguments elaborated in (Hladká et al., 2011), we have postponed the transformation of the syntactic annotations of the CAC spoken texts. Nevertheless, we are considering the spoken text reconstruction implemented in the annotation of the Prague Dependency Treebank of Spoken Czech (Hajič et al., 2017). This strategy is based on the fact that on the PDT a-layer it is not allowed to add, delete, and merge nodes, change spelling, move punctuation etc. Such reconstruction would make spoken text closer to written grammatically correct and comprehensible text, and thus the instructions for written texts can be used to annotate spoken texts. Example (3) illustrates the reconstruction of Example (2), the removed words are given in brackets.

The CLARIN infrastructure reports that out of 63 spoken text corpora being available in its repositories only two of them contain syntactic annotations, namely the already mentioned Prague Dependency Treebank of Spoken Czech consisting of 120,000 words completely syntactically annotated and the Estonian Dialect Corpus with annotated 40,000 out of 1.3 million words.[4]

(2) *I profesoru Backvisovi to$_3$ bylo trapné, říkal, že jistě tedy ta čeština by si zasloužila místo nejméně jako ta polština, že, no tak, že mně to jako, když mluví on o té$_1$, jako polonista, o tom$_1$, no, ale nemohl to oddiskutovat, prostě říkali, že je to tedy$_2$, ty otázky, tedy$_2$ skutečně inflace, a ta hospodářská krize doléhá už právě nebo především také tedy$_2$ na vědu a školství, tedy.*

Lit. *Professor Backvis was embarrassed by this$_3$, he said, that surely thus the Czech would deserve the position at least as the Polish, that, well, that to me this as, when he is speaking about the$_1$, as a Polonist, about the$_1$, well, but he could not discuss it, they just said, that this is thus$_2$, those questions, thus$_2$ indeed economic crisis already influences right now or above all also thus$_2$ on science and education, thus$_2$.*

(3) *I profesoru Backvisovi to bylo trapné, říkal, že jistě (tedy ta) čeština by si zasloužila místo nejméně jako (ta) polština, (že, no tak,) že (mně to jako,) když mluví on (o té$_1$,) jako polonista (,) o tom$_1$, (no,) ale nemohl to oddiskutovat, prostě říkal(i), že (je to tedy$_2$,) ty otázky, tedy$_2$ (skutečně) inflace(,) a (ta) hospodářská krize doléhá už právě nebo především také (tedy$_2$) na vědu a školství (, tedy).*

Contribution

We studied whether the morphological and syntactic annotations of the thirty-year-old Czech Academic Corpus can be automatically transformed into the scheme of Prague Dependency Treebank which has been dominating the annotation projects on Czech since 1990s. Partly we carried out the automatic transformation and partly the manual annotation. We also studied the details of the spoken text syntactic annotation, namely which annotation instructions formulated for written texts can be applied to spoken texts.

[4] https://www.clarin.eu/resource-families/spoken-corpora

2.2 Corpus-based Exercise Book of Czech

> **We think students need to learn the sentence structure. How to explore annotated corpora in language classes where morphology and syntax are taught and practiced using sentence diagrams?**

Students cannot practice morphology and syntax with the annotated corpora mainly because of the way in which the corpora are presented. Therefore we created the corpus-based exercise book of Czech morphology and syntax `Styx` that is compiled from `PDT 2.0` (Hladká and Kučera, 2005), (Kríž and Hladká, 2008). This task is analogous to the transformation of `CAC` so that the annotations in `PDT 2.0` are transformed into a target annotation scheme having the form of sentence diagrams taught at Czech schools. In this annotation scheme, a sentence is represented as a tree-like structure having no root node or, in another approach having two roots: a subject and a predicate, see (*sečíst, bylo nemožné*) in Figure 3. In contrast with the `PDT` annotation scheme illustrated in Figure 2, there is not a 1:1 correspondence between the number of nodes and tokens, e.g., the tokens *bylo, nemožné* as well as *ve, sněmovně* form single nodes where the tokens are listed according to their surface ordering in the sentence.

In general, exercise books are used to practice. Typically they contain true answers so that students can immediately check their answers. For practicing (or training) school diagramming, an exercise book should contain sentences with their analyses. To build an exercise book of a significant volume is a very demanding activity. However, if a corpus with relevant language phenomena annotated is available, an exercise book can be created (semi-)automatically from the corpus. We focused on the sentences in `PDT 2.0` that are annotated both morphologically and syntactically. However, there emerge some syntactic phenomena that are handled differently in the `PDT` annotation scheme than in the sentence school diagrams. Given that, the sentences annotated also on the t-layer were taken into account to process these phenomena properly with respect to the school approach. There are 49,442 such sentences in `PDT 2.0`. However, not all of them were included into the exercise book, mainly because of their complexity. We formulated several filters to exclude problematic sentences from the set of candidates. In the end, we got 11,718 sentences that we included into our exercise book and we transformed their `PDT 2.0` annotations into the diagrams by the rule-based transformation. While the transformation of the morphological annotations was straightforward, the transformation of the syntactic annotations was more complex. We defined the three key operations on the `PDT` dependency trees and mapping rules for the `PDT` analytical functions. Then the syntactic transformation is a sequence of key operations and mapping rules. The transformation of the `PDT` tree in Figure 2 into the diagram in Figure 3 applies the `AbsorbTheChildNodes` rule twice, namely on the nodes *bylo*, *nemožné* and *ve*, *sněmovně*.

The `Styx` exercise book contains not only sentences and their analyses but software systems as well. Manual browsing of more than 11 thousand sentences is impossible; the `Charon` system enables to select sentences for practicing according to the pre-defined criteria, e.g., the criterion *Sentences with subject* filters out the sentences without subject (Czech is a pro-drop language so subjects can be missing in sentences). Consequently, students draw diagrams for the selected sentences using the `Styx` system. We presented `Styx` and the methodology how to work with `Styx` to both students and teachers (Veselá, 2012).[5]

[5] `http://ufal.mff.cuni.cz/styx`

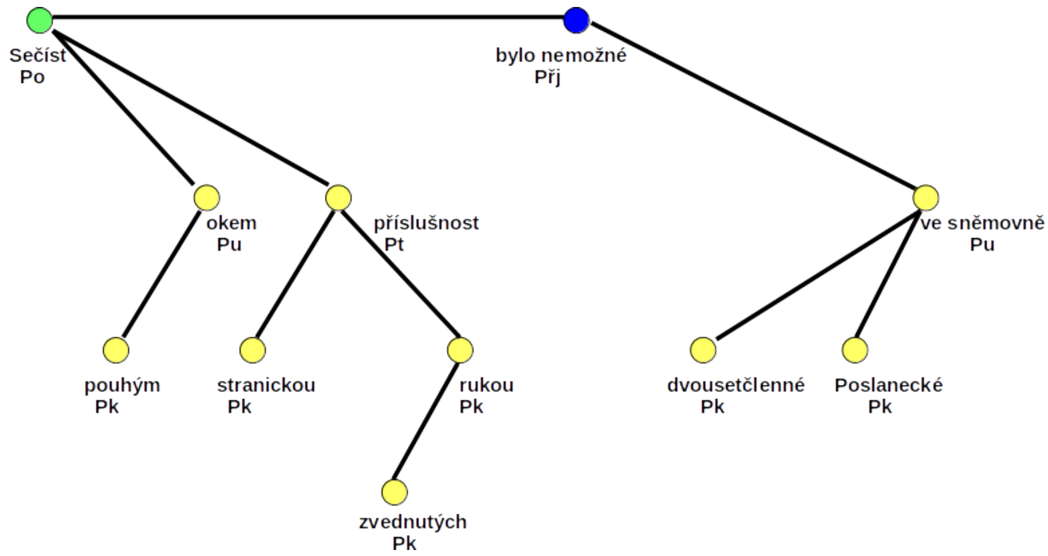

Figure 3: Sentence diagram of Example (1).

The sentence diagrams have been available only while running `Styx` and thus their linguistic analysis was impossible. It motivated us to apply the transformation operations and rules outside `Styx` and it resulted in the annotated corpus `STYX 1.0` (Hladká et al., 2017).

Contribution

We used the `Prague Dependency Treebank` of enormous volume to create the corpus-based exercise book of Czech morphology and syntax of enormous volume as well. We selected from `PDT` the sentences suitable for students and using the rule-based transformation we created their sentence diagrams. We implemented the `Styx` editor for students and teachers to select sentences, draw their sentence diagrams, and check them automatically. The sentence diagrams are available in the `STYX 1.0` corpus.

2.3 Czech Legal Text Treebank

> **We need gold standard corpus to test the information extraction system that explores a parsing procedure. How much of the experience acquired from the annotation of newspaper texts is applicable to the annotation of legal texts?**

We created the `Czech Legal Text Treebank` to evaluate the `RExtractor` system performance. `RExtractor` implements an extraction pipeline which processes input texts by linguistically-aware tools and extracts entities and relations using queries over dependency trees (see Chapter 4). The sentences in `CLTT` are taken from the Collection of Laws of the Czech Republic, namely The Accounting Act, 563/1991 Coll. and The Decree on Double-entry Accounting for Undertakers, 500/2002 Coll. The m- and a-layers of the `PDT` annotation scheme were used as the target annotation scheme and we enriched the a-layer annotation with the entity and relation annotations, for illustration see Figure 6.

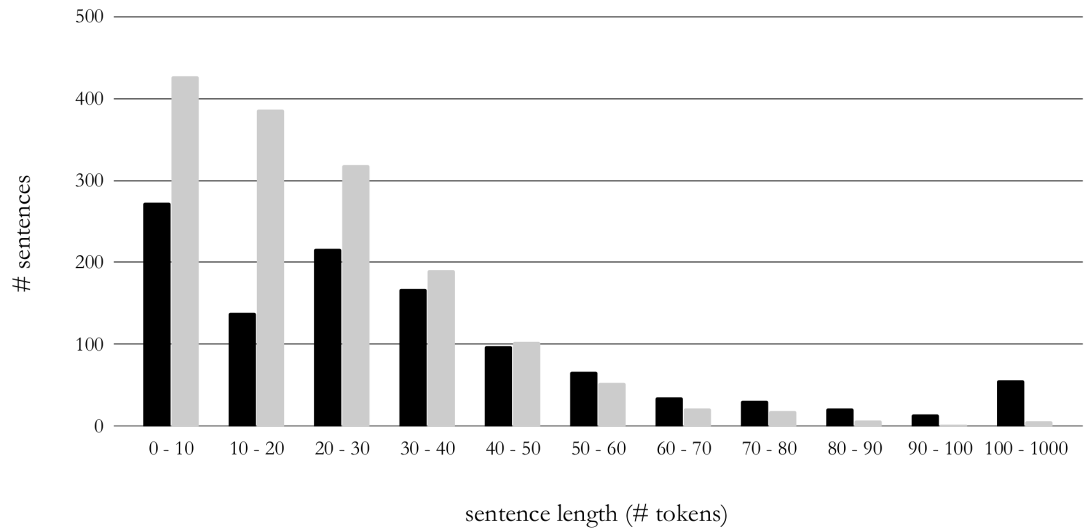

Figure 4: The sentence length distribution in `CLTT 2.0` when the `Treex` sentence segmentation and tokenization procedures are used (dark bars) and when their output is processed by the re-segmentation and re-tokenization rules (light bars).

Morphological and syntactic annotation The texts were tokenized and segmented into sentences. The sentence segmentation and tokenization procedures implemented in the `Treex` framework split the `CLTT` texts into 1,221 sentences having the length distribution displayed in Figure 4. Both long sentences and complex tokens make manual annotation very difficult. Also, a parsing procedure shows lower performance on long sentences, as evidenced in Chapter 4. Therefore we manually designed the rules to split long sentences into shorter segments and to join tokens into complex ones. Figure 4 visualizes their influence on the sentence length and Figure 7 illustrates re-tokenized tokens, see the node § *1 odst. 2 písm. d) až h)*. Afterwords `CLTT` was automatically morphologically and syntactically annotated using the `MST` parser (McDonald et al., 2005) adapted for Czech by Novák (2007). Then both the dependency trees and the analytical functions were manually checked. The annotator first checked each sentence segment individually and then used inter-segment links to capture dependencies between nodes from different segments. In order to do both the manual correction of partial segment dependency trees and the inter-segment linking we exploited the tree editor `TrEd` complemented by the extension `Czech Legal Text Treebank 2.0`.[6]

Entity annotation In cooperation with the experts of Sysnet, Ltd., one of the partners of the INTLIB project, we manually detected and classified domain-specific entities in the `CLTT` texts using `Brat` editor.[7] Consequently, the `Brat` entity annotations were manually annotated in the dependency trees, for illustration see the entities *účetní jednotka* (*accounting entity*), *majetek* (*asset*), and *závazek* (*liability*) highlighted in Figure 6. At the same time, these annotations were automatically transformed into the queries formulated in the PML-Tree Query language (Pajas and Štěpánek, 2006). A PML-TQ query extracts nodes from dependency trees that satisfy given requirements on their properties. The PML-TQ query in Figure 5 shows how to search the entity

[6] `https://ufal.mff.cuni.cz/tred/`
[7] `http://brat.nlplab.org/`

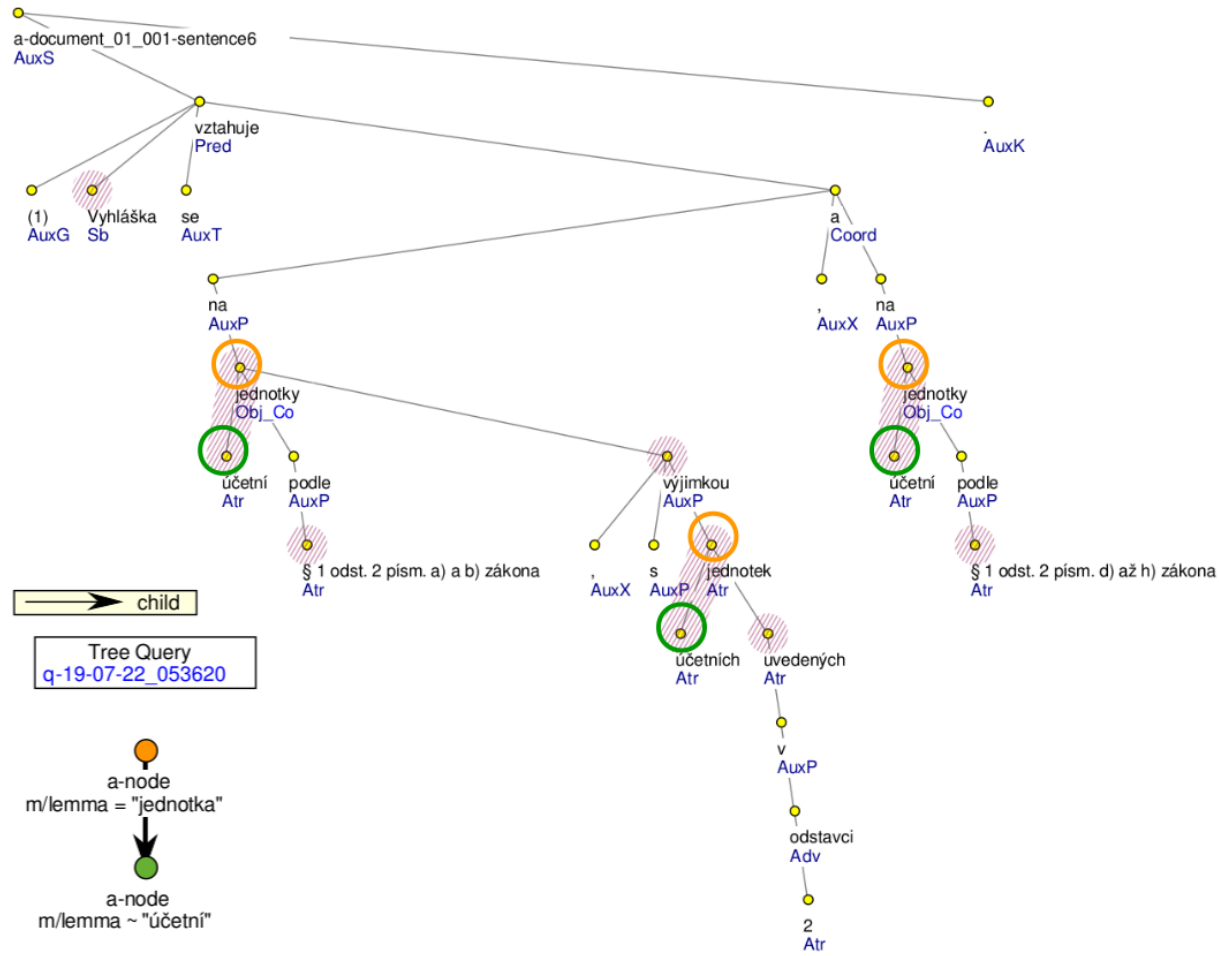

Figure 5: A sample PML-TQ query to search the entity *účetní jednotka* (*accounting unit*). There are three results of the query in the dependency tree.

účetní jednotka (*accounting unit*). The query defines two nodes – a daughter node with the lemma *jednotka* (*unit*) depends on its mother node with the lemma *účetní* (*accounting*).

Relation annotation A relation between entities is defined as a triple (*subject*, *predicate*, *object*) where *subject* and *object* are entities and *predicate* is a lexical representation of the relation. Three types of relations were annotated manually:[8]

- `obligation` where *subject* has an obligation to do *object* for legal reasons, see the relations (*accounting units*, *take inventory*, *assets*) and (*accounting units*, *take inventory*, *liabilities*) in Example (4)

> (4) *Účetní jednotky jsou povinny inventarizovat majetek a závazky podle § 29 a 30.*
>
> Lit. *Accounting units shall take inventory of their assets and liabilities pursuant to section 29 and 30*

[8]English translations of Examples (4), (5), and (6) are taken from `https://tinyurl.com/y6f8e7r4`.

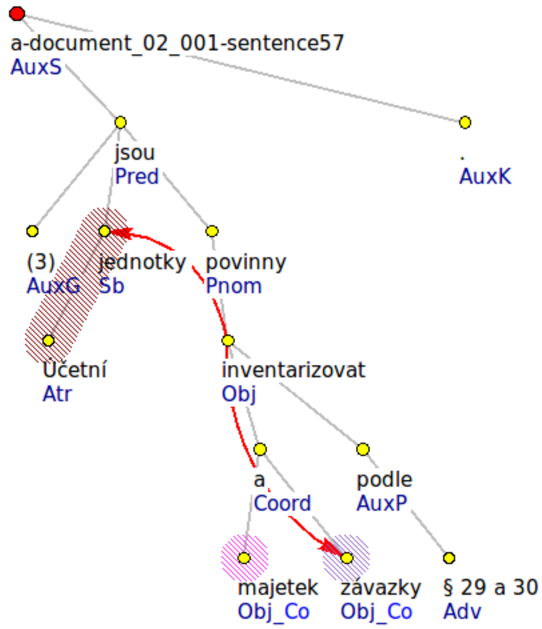

Figure 6: Example (4) in `CLTT 2.0`: three entities and one `obligation` relation highlighted.

- `right` where *subject* has a right to do *object* for legal reasons, see the relation (*accounting units*, *terminate*, *keep accounting*) in Example (5)

  (5) *(7) S výjimkou ukončení činnosti mohou účetní jednotky podle §1 odst. 2 písm. d) až h) ukončit vedení účetnictví nejdříve po uplynutí 5 po sobě jdoucích účetních období, ve kterých vedly účetnictví.*

  Lit. *(7) Except for the case of terminating activity, accounting units referred to in section 1(2)(d) to (h) may terminate to keep accounting earliest after expiry of five successive accounting periods in which they kept accounting.*

- `definition` where *subject* is defined as *object* representing a statement of what a term (*subject*) means, see the relation (*market value*, *mean*, *closing price*) in Example (6)

  (6) *Pokud je majetek veden na regulovaném trhu, rozumí se tržní hodnotou závěrečná cena vyhlášená na regulovaném trhu v pracovní den, ke kterému se ocenění provádí.*

  Lit. *If a particular asset is quoted on this country's stock exchange, the market value shall mean the closing price listed by the stock exchange on the business day when the valuation is made.*

We used the `Brat` editor for the relation annotation. The relations are visualized in dependency trees using oriented links going from the *predicate* node to the roots of subtrees representing *subject* and *object*, see Figures 6, 7, and 8. In Figure 6 only

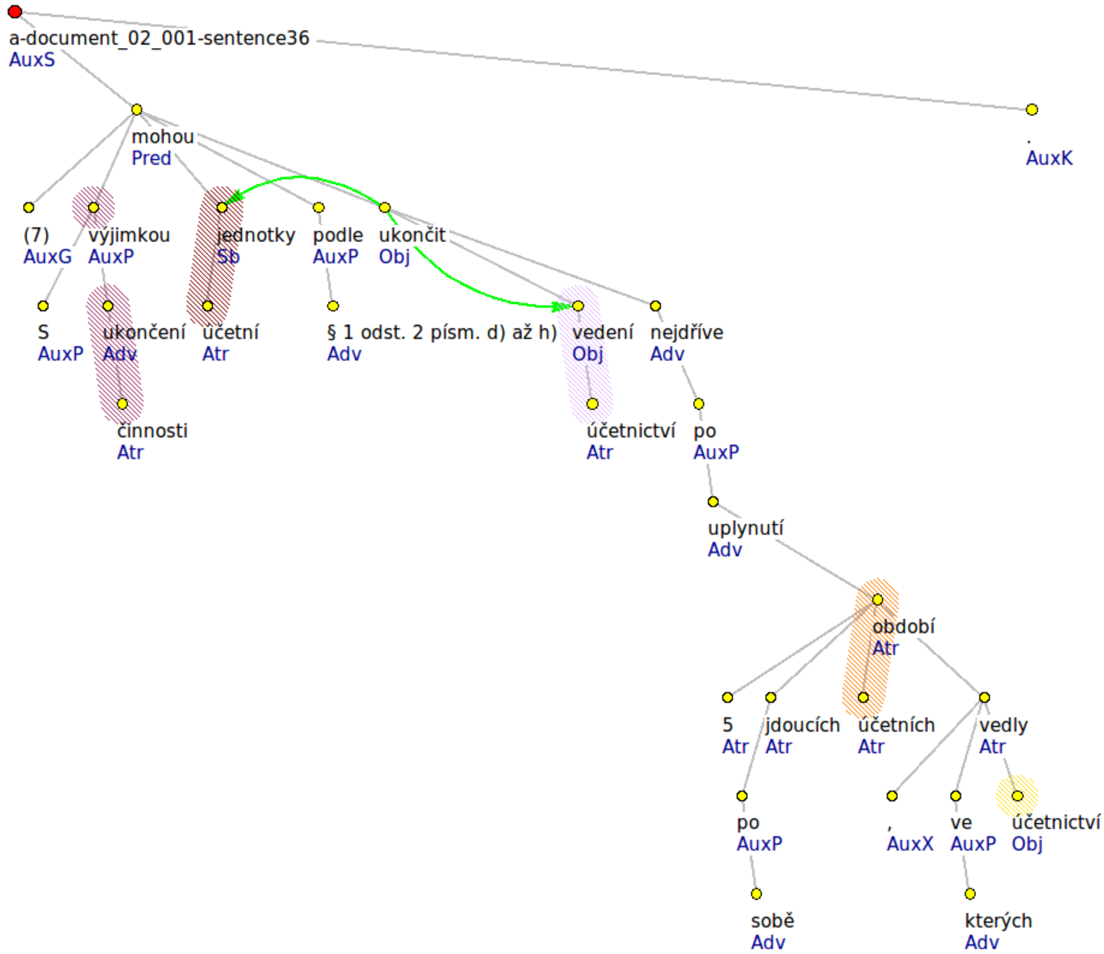

Figure 7: Example (5) in `CLTT 2.0`: six entities and one `right` relation highlighted.

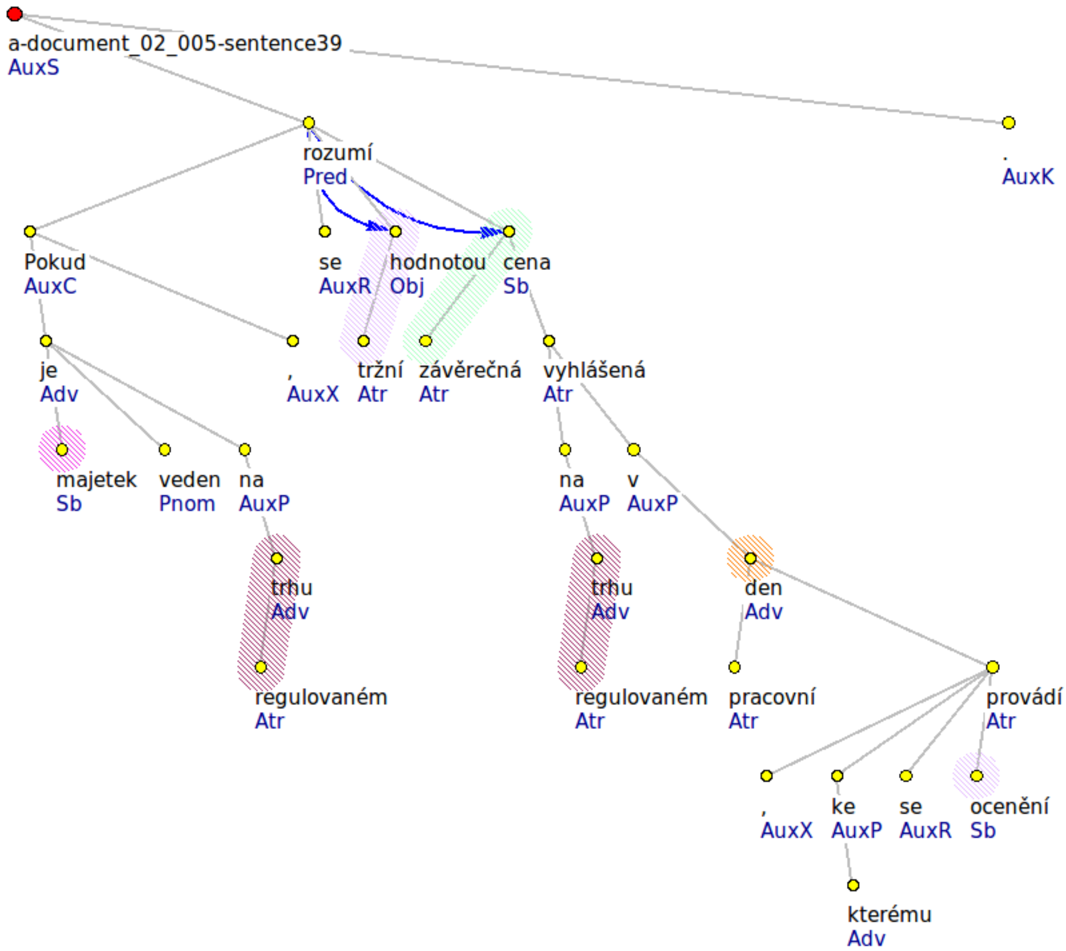

Figure 8: Example (6) in `CLTT 2.0`: seven entities and one `definition` relation highlighted.

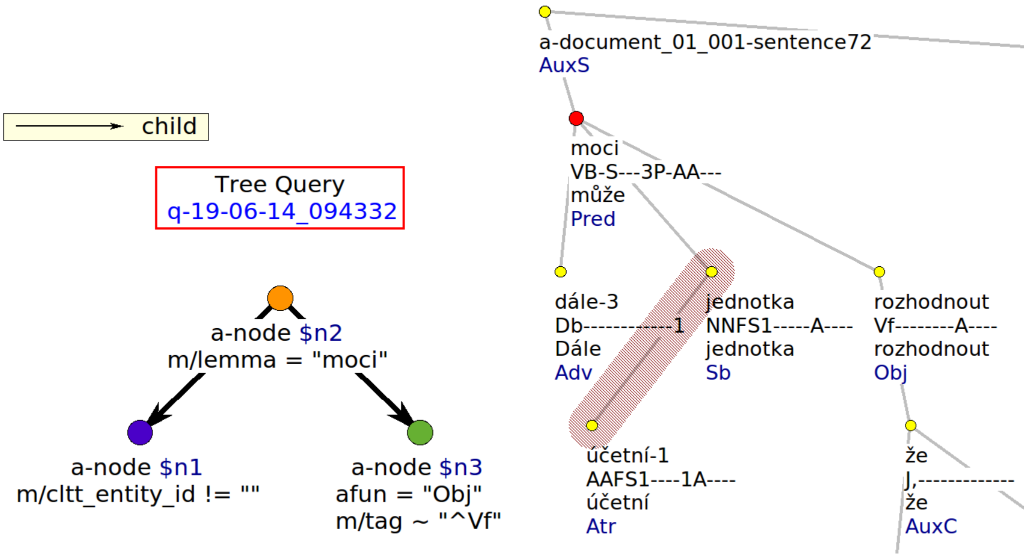

Figure 9: The `obligation` relation (*účetní jednotka/accounting unit*, *rozhodnout/decide*, ...) in the tree on the right matches the PML-TQ query on the left that extracts an obligation (*moci/may*, `a-node$n2`) of an entity (`a-node$n1`) to do (`a-node$n3`) something.

the `obligation` relation (*accounting units*, *take inventory*, *liabilities*) is visualized, although in total two relations are in the sentence the *objects* of which are members of the coordination *assets and liabilities*. This way of visualization is our convention and a proper tracing of the dependency tree detects the relation (*accounting units*, *take inventory*, *assets*) as well.

Like the entity annotations, the relation annotations were transformed into the PML-TQ queries. In Figure 9 the query for searching the `obligation` relations is displayed together with the dependency tree that matches this query: the predicate has the lemma *moci* (*may*) and is connected with an entity (*accounting unit*) and an infinitive (`Vf*`) *rozhodnout* (*decide*) bearing the analytical function `Obj`(ject).[9]

Contribution

We carried out the morphological and syntactic annotation of the `Czech Legal Text Treebank` using the guidelines formulated for the `Prague Dependency Treebank`, an annotated corpus of the newspaper domain. We paid a special attention to complex tokens and sentences that occur more frequently in legal texts than in newspaper texts. To make the manual annotation of dependency trees as organized as possible, we modified both the `PDT` annotation strategy and the `TrEd` editor.

[9]More queries are available on `http://lindat.mff.cuni.cz/services/pmltq`.

3. Alternative Annotation

We define an alternative way of corpus annotation as a crowdsourcing procedure to increase the volume of annotated data by decreasing total costs of academic annotation projects.

Crowdsourcing is the practice to engage a large number of people, a crowd, to meet a common goal. No doubt the Internet and schools are the most appropriate environments to implement crowdsourcing projects. Mainly because the Internet accelerates communication, concentrates human resources, enables information sharing, and supports volunteering. On the other hand, schools provide space to students to learn and practice. However, it is still desirable to carry out academic annotation in parallel mainly because (1) a crowdsourcing project can attract only few users who produce not enough data to improve performance of a given task, and (2) academic annotations can be used to measure quality of annotations produced by users and to test their reliability. We focused on two alternative annotation strategies that use on-line games and school sentence diagramming. Table 2 provides their stories in a nutshell.

annotation	academic	alternative	alternative
title	`CAC`, `CLTT`, `STYX`	`PlayCoref`	`Čapek`
format	research project	game	language class
team	linguists	players	students
required knowledge	annotation framework	language knowledge	language grammar
instructions	annotation guidelines	game rules	language grammar
environment	editor	game	editor
gold annotation	arbiter	agreement	combination

Table 2: Academic and alternative annotation in a nutshell.

3.1 Play the Language

> **How to use enjoyment of crowds of game players to get more gold standard data for textual procedures that have not achieved human performance yet and thus to decrease total costs of academic annotation projects?**

Luis von Ahn, a pioneer in the field of human computation, invented on-line Games With A Purpose (GWAPs) that are designed to annotate data for those tasks that have not achieved human performance yet. In addition, users play GWAPs and produce annotated data as a by-product of enjoyment (von Ahn, 2006), (von Ahn and Dabbish, 2008). The very first game that his team implemented was the `ESP` game where an image is shown to two players and they label it with words they expect the opponent will use (von Ahn, 2006).[1] Then they implemented the two-player game `TagATune` with audio: each player listens to a short excerpt of music, one player describes it with words and guesses whether he listens to the identical audio as his opponent (Law et al., 2007).[2] Their `Verbosity` game is a two player game as well: a player-narrator thinks a word and using the pre-formulated parts of sentences he offers help to his opponent who guesses his word (von Ahn et al., 2006). Like in the academic annotation, also GWAPs require elements to ensure the quality of annotations. The most important elements are the agreement reached by a significant number of players, the meaningful annotations, and the reliability of players. In the `ESP` game, images are being displayed in more than one session and players cannot use the words from a stop word list.

The GWAPs have met with a great success, especially the `ESP` game with which the users labeled incredible amount of images in a short time, as evidenced by e.g., "*as of July 2008, 200,000 players had contributed more than 50 million labels*" in (von Ahn and Dabbish, 2008). No wonder they inspired other researchers from various fields of study as illustrated in the survey by Lafourcade et al. (2015). We can see there that the GWAPs with images were created first, followed by the games with sound recordings and finally with video recordings and texts. This is not a coincidence. This fact is strongly related to the amount of time it takes players to comprehend what they are watching/listening to/reading: while images require a short time, texts are more time consuming and playing the games with texts tends to be unexciting. As evidenced by (Chamberlain et al., 2017), (Chamberlain et al., 2018), the games in NLP are the most recent ones with two exceptions, our `PlayCoref` game on coreference resolution and the `Phrase Detectives` game on anaphora resolution (Chamberlain et al., 2008), (Poesio et al., 2013). The survey by Sukthanker et al. (2018) presents an exhaustive overview of the anaphora and coreference resolution field. None of the presented systems achieved the F1 score higher than 0.8. For Czech the best automatic coreference resolution system shows the F1 score below 0.7 (Novák, 2018). The manual coreference resolution shows substantially higher performance and this motivated us to design a GWAP for the coreference resolution task, the `PlayCoref` game. In addition, we developed two more games `Shannon game` and `Place the Space`. The three GWAPs with Czech and English textual data are available at our `LGame` portal.[3] In their design, we relied on our knowledge of NLP, both data and tools.

The games `Shannon game` and `Place the Space` were implemented mainly to attract users. `Shannon game` is a single-player and two-player game on reconstruction of the hidden words in the input sentences. `Place the Space` is a single-player game on reconstruction of the hidden spaces in the input sentences.

(7) *One Flew Over the Cuckoo's Nest is a 1975 American comedy-drama film*$_3$ *directed by Miloš Forman*$_1$*, based on the 1962 novel One Flew Over the Cuckoo's Nest by Ken Kesey. The film*$_3$ *stars Jack Nicholson*$_2$ *as Randle McMurphy ...As Forman*$_1$ *did not allow the actors to see the day's filming, this led to the cast losing confidence in him*$_1$*, while Nicholson*$_2$ *also began to wonder about his*$_2$ *performance. Douglas convinced Forman*$_1$ *to show Nicholson*$_2$ *something*$_4$*, which*$_4$ *he*$_1$ *did, and restored the actor's*$_2$ *confidence ...* (Source: Wikipedia)

The `PlayCoref` game is a single-player and two-player game on coreference, a linguistic phenomenon that crosses sentence boundaries. Coreference describes the relation among two or more expressions that refer to the same discourse entity in the text.

[1] `http://en.wikipedia.org/wiki/ESP_game`
[2] `http://tagatune.org/`
[3] `http://ufal.mff.cuni.cz/tools/lgame`

The subscripts in Example (7) distinguish the objects they refer to, e.g., the expressions $Forman_1$, $Forman_1$, him_1, he_1 refer to the entity (person) of Miloš Forman.

In `PlayCoref` the players read a short text and connect words that co-refer. Their task is to connect words in as many sentences as possible and the sentences of the texts are displayed into sessions on player's requests. The `Phrase Detectives` game on anaphora resolution was designed in parallel with `PlayCoref` and the annotations created while playing the game were published as the part of `Phrase Detectives Corpus 1.0` eight years later (Chamberlain et al., 2016).[4] The anaphora is a relation that points back to an expression in the text, e.g., *Forman* and *him*. The main difference between the two games is in their basic principles: `Phrase Detectives` game offers the player a whole paragraph and asks him one specific question at a time, e.g., "Does this word co-refer with another word in the previous text? If so, with which one?" `PlayCoref`, on the other hand, presents the text to the player sentence by sentence and asks him to search for all coreferential relations.

We highlight those `PlayCoref` features that make it novel: **(1)** input data preprocessing – we want to use as many NLP procedures as possible. We therefore run the procedures of sentence segmentation, tokenization, tagging, and coreference resolution on the input texts; **(2)** word locking – among other things, the quality of the game data is influenced by what actions the players can do. It is desirable to appropriately navigate them to correct or at least meaningful annotations. In `PlayCoref` such navigation takes the form of locking the words in the text that cannot co-refer. Subsequently, it is possible to play only with the unlocked parts. Technically, we lock the morphologically tagged words according to the properties of grammatical and textual coreference. For illustration see the struckthrough words in the following sentence *Douglas ~~convinced~~ Forman ~~to~~ ~~show~~ Nicholson something ~~,~~ which he ~~did~~ ~~,~~ ~~and~~ ~~restored~~ ~~the~~ actor's confidence ~~.~~* **(3)** gold data – if a corpus manually annotated with coreference is available for the language of the game, it can be used to evaluate the annotations coming from the game sessions. The `Prague Depedency Treebank` is the only corpus where coreference is annotated on the underlying dependency-based syntactic layer (t-layer), see e.g., (Zikánová et al., 2015). Therefore we mapped the annotations from the trees to the surface layer. The coreference annotation in other corpora, e.g., `Polish Coreference Corpus, MUC-6, BBN Pronoun Coreference and Entity Type Corpus` was carried out on the surface layer;[5] **(4)** player scoring – we reward players A and B for the coreferential links that they made in s sentences that they read using the formula $\lambda_1\times$ F(A, acr or gold) + $\lambda_2\times$ F(A, B) + $\lambda_3\times$ $\min(12, s)/12$, where F is the F1 score, 'acr' stands for an automatic coreference resolution procedure and 'gold' for a gold annotation. We motivate players to read at least twelve sentences. The weights for the sub-scores $\lambda_1, \lambda_2, \lambda_3$ are set empirically.

Contribution

The present procedures of automatic coreference resolution have not achieved human performance yet. To enlarge the gold data for these procedures, we designed and implemented the `PlayCoref` game, a game with a purpose producing annotations as a by-product of enjoyment. We focused on the design of game features to get annotations of as high-quality as possible.

[4] http://www.phrasedetectives.org

[5] http://core.ipipan.waw.pl, http://catalog.ldc.upenn.edu/LDC2003T13, http://catalog.ldc.upenn.edu/LDC2005T33

3.2 Sentence Diagramming

Can we engage students learning and practicing morphology and syntax using sentence diagrams into corpus annotation so that we collect their sentence diagrams and transform them into another annotation scheme?

We consider language classes at schools as another place where we can search for alternative annotators. We designed a crowdsourcing way of the involvement of students into creating a morphologically and syntactically annotated corpus in three steps (1) collection of sentence diagrams created at language classes, (2) combination of the diagrams into single better diagrams, (3) transformation of the final diagrams into a target academic annotation scheme.

To perform (1) a tool-editor for drawing and collecting sentence diagrams has to be available. We have been developing the client-server application `Čapek` that we have named after Czech writer Karel Čapek and we have designed it both as a crowdsourcing system for getting annotated data and a Computer-Assisted Language Learning system for practicing morphology and dependency-based syntax (Hana and Hladká, 2012). To get diagrams of high quality, we prefer to have more than one diagram for each sentence in the pool of sentences analyzed by students. If so, we can proceed to the step (2): we formulated the measure of tree-edit-distance to quantify similarity of two diagrams for the same sentence as the minimal cost of turning one into another using a set of simple operations (the smaller the distance, the similar the diagrams). Second, we designed a combination procedure based on majority voting that first determines the set of nodes and then the set of edges over these nodes (Hana et al., 2014).

In the pilot study, we randomly selected a sample of 100 sentences from Czech textbooks and we organized their sentence diagram annotation by two teachers, two secondary school students, and seven undergraduates using `Čapek` (Konárová, 2012). We designed a rule-based transformation of diagrams into the `PDT` annotation scheme that is reverse to the one employed in the building of the `Styx` exercise book.

We think students should be trained in understanding sentence structure and thus we cannot see their effort to draw diagrams as wasting their time. From the curriculum perspective it is also very interesting to see the differences in teachers' and students' diagrams. In this study, we obtained mixed results, with relatively low agreement between the teachers, and high agreement between the students. However, if we want to get more annotated data via the school annotation and its transformation, a parser's accuracy must be beaten. In this study, the accuracy of the proposed procedure is lower than the `MST` parser's accuracy. Since we have the annotations by a limited number of users and rules, we cannot make final conclusions. We are aware of phenomena not covered by these rules and we believe that covering them will significantly improve the transformation accuracy.

Contribution

We implemented the `Čapek` editor by means of which students and teachers can draw sentence diagrams electronically and upload them to our remote repository. We developed a metric to quantify differences between diagrams and formulated a procedure to combine them into a single one. We conducted a pilot study with diagrams produced in Czech language classes being transformed into the `Prague Dependency Treebank` annotation.

4. Information Extraction

Jurafsky and Martin (2009) describe Information Extraction (IE) as a process that turns unstructured information embedded in texts into structured data. Most IE tasks start with the task of Named Entity Recognition where each mention of a named entity is first recognized and then classified into pre-defined categories such as the person names and events, e.g., *Miloš Forman*, *Ken Kesey*, *Academy Award* in Example (8). Entities recognized in a text can be linked together to refer to real-world entities. This is the task of coreference resolution that we mentioned in Chapter 3 on the alternative annotation. Once the entities are recognized IE tasks typically continue with the task of recognition and classification of semantic relations among them. These relations are represented as n-ary relations, e.g., the binary relation *director (Miloš Forman, One Flew Over the Cuckoo's Nest)* and the 3-ary relation *novel (Ken Kesey, One Flew Over the Cuckoo's Nest, 1962).*

(8) *One Flew Over the Cuckoo's Nest is a 1975 American comedy-drama film directed by Miloš Forman, based on the 1962 novel One Flew Over the Cuckoo's Nest by Ken Kesey ... The film was the second to win all five major Academy Awards (Best Picture, Actor in Lead Role, Actress in Lead Role, Director and Screenplay) ... As Forman did not allow the actors ...* (Source: Wikipedia)

A collection of human-written unstructured documents related to the legal and environmental domains was assumed on the input of the INTLIB applied research project and its goal was to process the data in the extraction and presentation phase. The INTLIB team designed the extraction phase as a two-stage procedure where (1) entities and relations are extracted from the documents, and (2) they are represented in the framework of Linked Open Data. The presentation phase provides an efficient and user friendly visualization and browsing the extracted knowledge. The NLP group of the team addressed the task (1) respecting scientific aspects of the design, implementation, and evaluation of IE systems.

Both the legal and the environmental domains are still largely underrepresented in the Natural Language Processing literature despite its potential for generating interesting research questions. This fact and a multi-disciplinary nature of INTLIB make our work original. Table 3 provides the story of the two systems that we developed in a nutshell.

4.1 Legal Domain

> **Our aim is to use linguistic procedures in a system of information extraction of entities and relations from acts. Is the extraction from dependency trees instead of unstructured texts of benefit to the system?**

We designed and implemented `RExtractor`, a system for the entity and relation extraction by querying dependency trees (Kríž and Hladká, 2015). Figure 10 shows its general workflow designed independently on the domain and language under consideration. The input document first enters the technical Data Format Conversion component to convert its format. Then the Natural Language Processing component generates for

tool	`RExtractor`	`EIA extractor`
domain	legal	environmental
documents	Acts	EIA documents
platform	`Treex`	`Gate`
language	Czech	Czech
tagger	`MorphoDita`	`MorphoDita`
parser	`MST`	×
query language	`PML-TQ`	regular expressions
entity queries	semi-automatically	manually
relation queries	manually	manually

Table 3: Information Extraction systems in a nutshell.

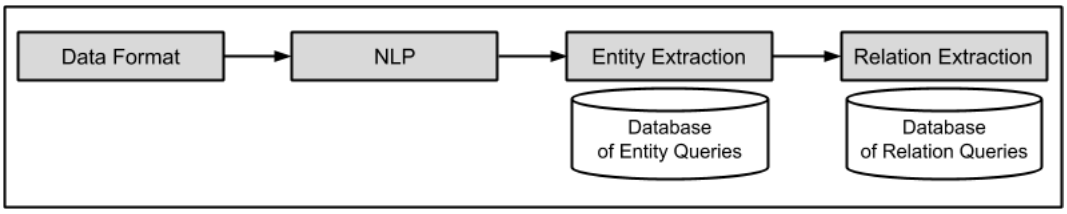

Figure 10: General workflow of Information Extraction systems.

each sentence of the document its dependency tree that is queried in the following two components. The Entity Extraction component detects in the trees the entities stored in the Database of Entity Queries and the Relation Extraction component detects the relations stored in the Database of Relation Queries. The queries are formulated in the `PML-Tree Query` language that we illustrated in Chapter 2 on the annotation of the `Czech legal Text Treebank`.

The NLP component is language dependent while the Entity Extraction and Relation Extraction components are domain dependent. We configured `RExtractor` for processing Czech documents accordingly: we use the NLP procedures trained on Czech gold corpora – the `MorphoDita` tagger and the `MST` parser, and filled the databases with the entity and relation annotations present in `CLTT 2.0`. Both databases can be enlarged or built incrementally so that more general queries are formulated as a modification of the specific ones.

Figure 11 illustrates the annotation of Example (9) in the `Czech Legal Text Treebank`. The entities are highlighted and the relation of `obligation` (*accounting units*, *account*, *business results*) is visualised using the arrows. In fact, there occur ten more relations in Example (9) that are not showed in the tree but that are extractable by tracing the dependency tree according to the analytical functions for coordination `*_Co`.[1] This feature clearly illustrates what the dependency tree search allows to extract

[1] (*accounting units*, *account*, *position of property*), (*accounting units*, *account*, *position of assets*), (*accounting units*, *account*, *position of commitments*), (*accounting units*, *account*, *position of liabilities*), (*accounting units*, *account*, *movements of property*), (*accounting units*, *account*, *movements of assets*), (*accounting units*, *account*, *movements of commitments*), (*accounting units*, *account*, *movements of liabilities*), (*accounting units*, *account*, *costs*), (*accounting units*, *account*, *revenues*)

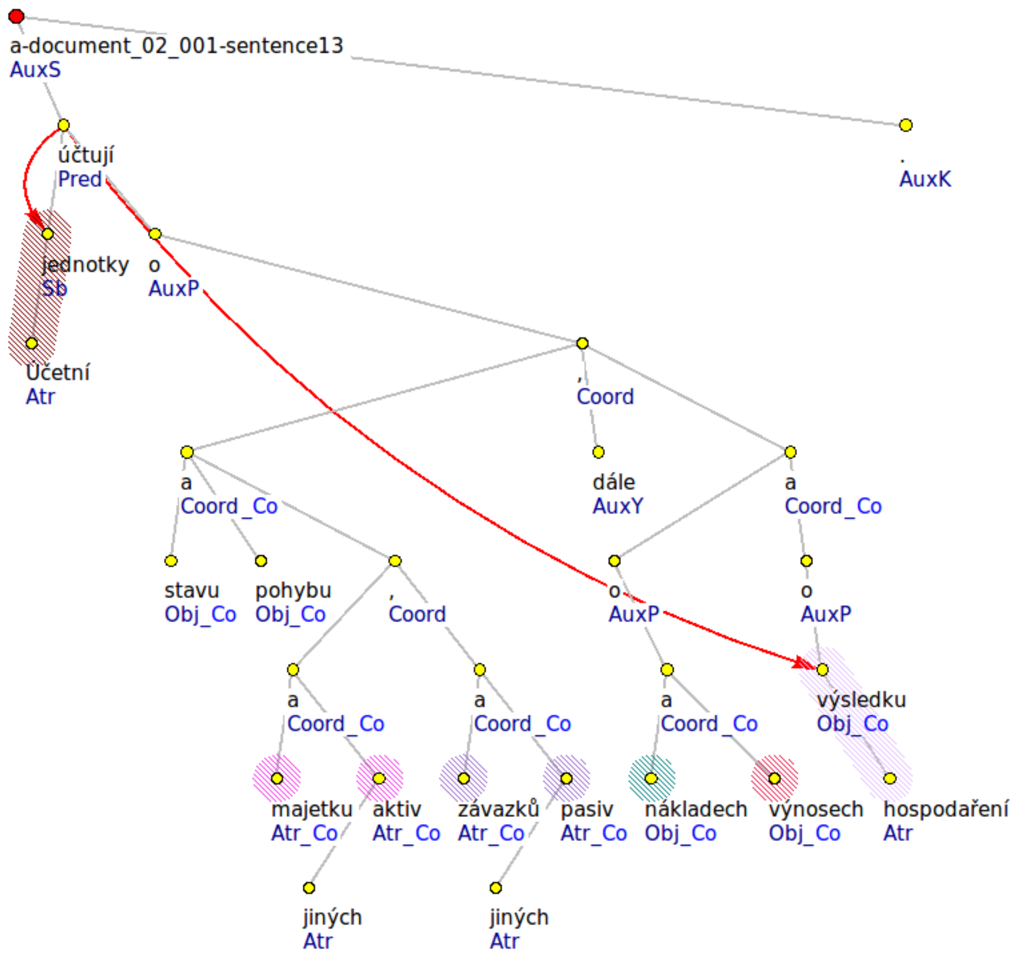

Figure 11: Annotation of Example (9) in `CLTT 2.0`.

beside the full-text search.

(9) *Účetní jednotky účtují o stavu a pohybu majetku a jiných aktiv, závazků a jiných pasiv, dále o nákladech a výnosech a o výsledku hospodaření.*

Lit. *Accounting units shall account for the position of, and movements in, their property and other assets, commitments and other liabilities as well as costs and revenues, and their business result.*

Example (10) and its dependency tree in Figure 12 illustrate another benefit of the dependency tree search: for the relation `right` (*accounting units*, *apply*, *economic year*) we go deep into the tree structure and we can see other attributes of the relation's *subject*, i.e., *other than organizational components of the state, self-governing territorial units and accounting units having been formed or established pursuant to a specific Act.*

(10) *(5) Účetní jednotky, které nejsou organizační složkou státu, územním samosprávným celkem nebo účetní jednotkou vzniklou nebo zřízenou zvláštním zákonem, mohou uplatnit hospodářský rok.*

Lit. *(5) Accounting units, other than organizational components of the state, self-governing territorial units and accounting units having been formed or established pursuant to a specific Act, may apply an economic year.*

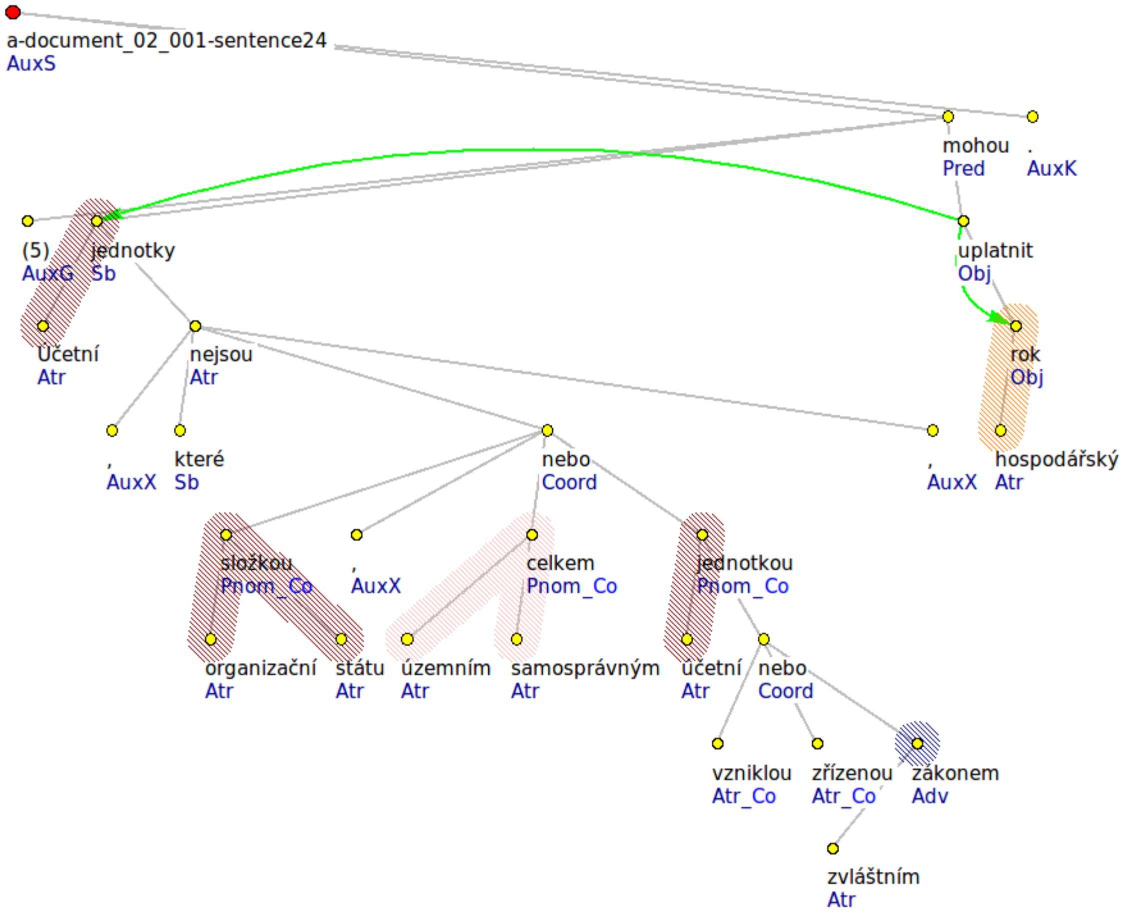

Figure 12: Annotation of Example (10) in CLTT 2.0.

Contribution

We built the RExtractor system that extracts domain-specific entities and the right, obligation, and definition relations from Czech acts. The extraction is implemented as a dependency tree matching method when input sentences are syntactically parsed first. This strategy allows to extract more data than full-text search especially those that occur in complex structures like coordination.

> **It has been empirically proven that syntactic parsing procedures show lower performance on complex sentences. Can we increase parsing performance by parsing clauses in sentences first and then linking their dependency trees to get final trees?**

RExtractor exploits dependency trees generated by a parsing procedure that influences its performance. We faced an interesting research question on cross-domain parsing since (i) a parsing procedure is the crucial part of the RExtractor extraction pipeline, and (ii) no parser trained on Czech legal documents exists. We overcame this obstacle by running the MST parser trained on the PDT 3.0 corpus from the newspaper domain. We tested its performance using the standard metric Unlabeled Attachment Score (UAS) that measures accuracy of dependencies in trees keeping dependency la-

treebank	dataset	# sentences	# tokens	`MST` UAS	`CCP` UAS
`CLTT 2.0`	orig	1,121	34,410	0.792	0.800
	multi	1,438	36,596	0.818	0.822
	train	29,768	518,648	0.934	×
`PDT 3.0`	dtest	4,042	70,974	0.845	0.850
	etest	4,672	80,923	0.843	0.849
`CAC 2.0`	written	24,709	493,306	0.827	0.836

Table 4: Unlabeled Attachment Score of the `MST` parser and the Clause Chart Parsing procedure (`CPP`).

bels (analytical functions) aside. Initially we were interested in changes in the `MST` parser performance for sentences of different lengths. We parsed `PDT 3.5` and `CAC 2.0` from the newspaper domain and `CLTT 2.0` from the legal domain; we parsed `CLTT 2.0` as is (`CLTT 2.0 orig`) and its 'multiplicated' version (`CLTT 2.0 multi`) illustrated in Examples (11) and (12) where a sentence with enumeration is split into simpler sentences. Figure 13 shows that UAS of `MST` decreases as the sentence length increases. For `CLTT 2.0` we see that its 'multiplication' apparently improves parsing of sentences longer than 30 tokens. Table 4 shows UAS of `MST` for the complete corpora.

(11) ***(1) Unless it is further provided for otherwise, accounting units shall open their books of account:***
(a) at the day when their obligation to keep accounting arises;
(b) at the first day of an accounting period;

(12) a. ***(1a) Unless it is further provided for otherwise, accounting units shall open their books of account*** *at the day when their obligation to keep accounting arises.*

b. ***(1b) Unless it is further provided for otherwise, accounting units shall open their books of account*** *at the first day of an accounting period.*

We focused on the idea of splitting the parsing process into parts mainly because of the high frequency of long sentences in legal documents. In NLP literature a few approaches deal with the idea of parsing split into several parts, e.g., chunk identification (Sang and Buchholz, 2000), cascade of specialized parsers (Ciravegna and Lavelli, 1999), shalow parsing strategies (Federici et al., 1996). Most recent approaches focus almost exclusively on improving full-scale parsing, see e.g. (Pei et al., 2015).

Initially we assume sentences in texts to be segmented into clauses. For Czech, we employed the rule-based clause segmentation procedure that operates over dependency trees (Bejček et al., 2013) and we can use clause charts to visualize its output, i.e. relationships between segmented clauses within the sentence and capture the layer of embedding of each individual clause. Technically, we represent a clause chart as a sequence of `B`'s and integers from the range $\langle 0, M\rangle$, where `B` stands for a clause boundary and M is the maximum layer of the chart. For illustration see the clause segmentation

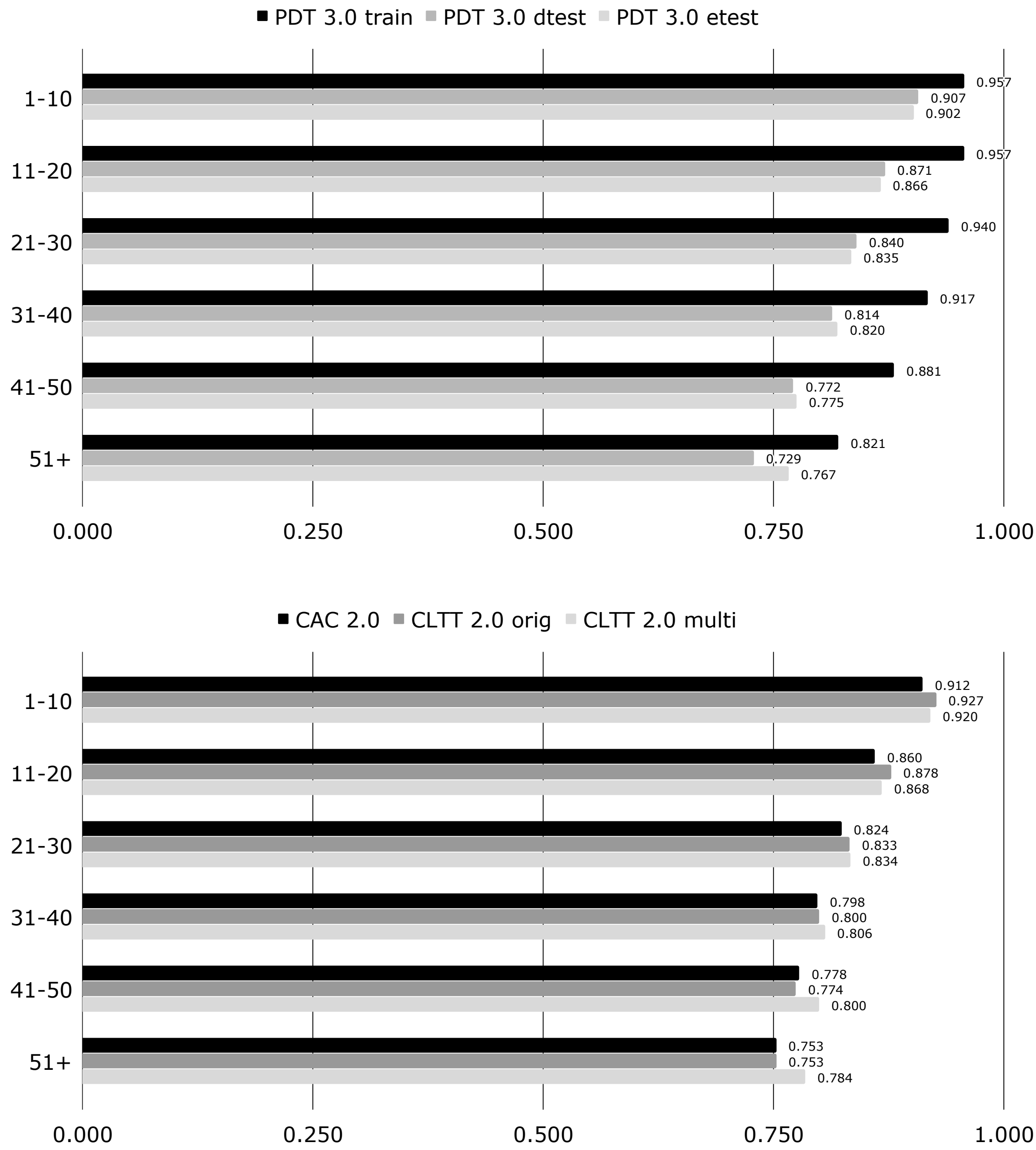

Figure 13: Unlabeled Attachment Score of the `MST` parser on the `PDT 3.0` and `CLTT 2.0` sentences of specific lengths. `MST` is trained on the `PDT 3.0` train data set.

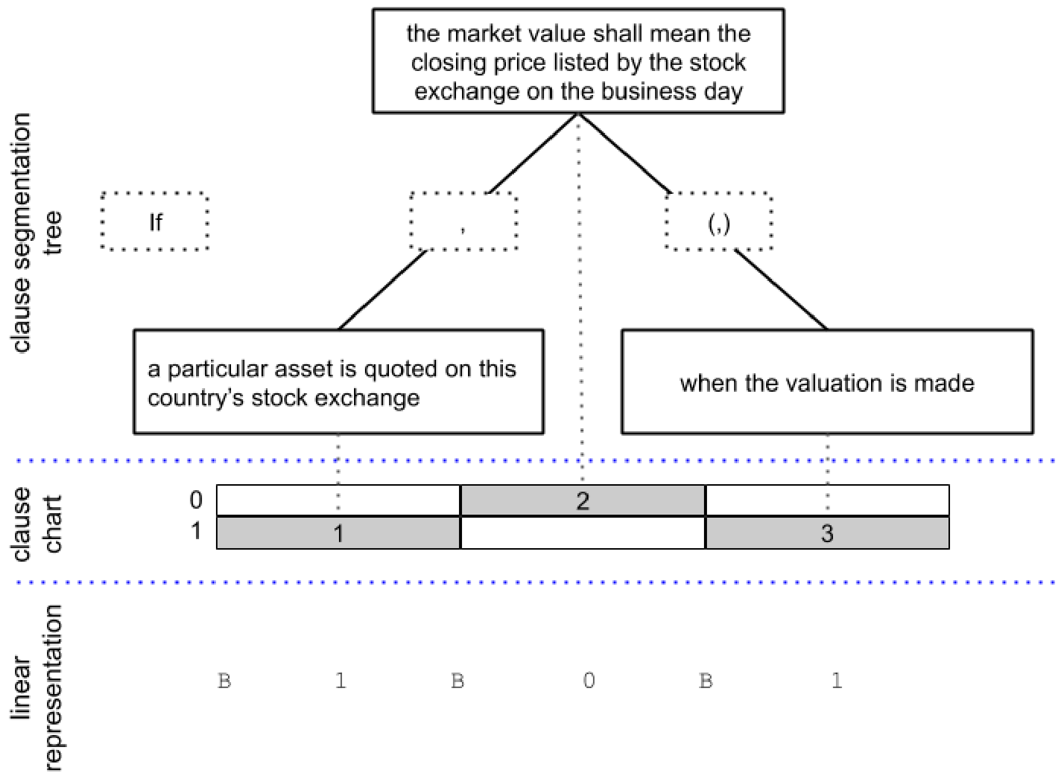

Figure 14: Clause chart of Example (6) and its linear representation

tree, clause chart and linear representation of Example (14) displayed in Figure 14. The other Examples listed in this thesis are represented as follows – Example (1) `0`, (4) `0`, (5) `0B1`, (9) `0`, (10) `0B1B0`.

We analysed the `MST` parser performance on the sentences of a particular chart pattern in great details and we focused on improving the parsing of sentences containing the `0B0` and `0B1` patterns, e.g., `0B0`, `0B1`, `0B1B0`, `0B0B0`. We formulated the Clause Chart Parsing procedure (`CCP`) and we achieved a higher performance in comparison to the `MST` parsing when sentences are parsed at once, see Table 4 (Kríž and Hladká, 2016). As our main goal was to build a knowledge base of entities and relations, we studied the influence which the parsing procedure has on the `RExtractor` performance. For the entity extraction, we used the PML-TQ queries stored in the Database of Entity Queries component to find the matching nodes in the dependency trees of the `CLTT 2.0` sentences. We experimented with the gold trees and the trees created by the `MST` parser and the `CCP` procedure and we reported the F1 scores 0.952, 0.937, and 0.946, resp. For the relation extraction, we achieved the F1 score 0.96 for the gold entities and 0.71 and 0.72 for the entities detected in the trees generated by the `MST` parser and the `CCP` procedure, resp.

Contribution

We formulated parsing of legal texts as a domain adaptation problem, i.e. we used the parser trained on the newspapers domain and applied it on the legal domain. We studied the complexity of long sentences using their clause charts and we designed a rule-based procedure to parse sentences by their clauses and then to link their dependency trees into a final tree. We achieved better results in comparison with the full-scale parsing.

4.2 Environmental Domain

Our aim is to use linguistic procedures in a system of information extraction of entities and relations from EIA documents on the environmental impacts of a given project. Is the extraction strategy implemented for acts applicable to EIA documents?

The most important feature of legal texts is their very specific syntactic structure with many peculiarities. We often encounter passive voice structures, impersonal constructions, non-finite and verbless clauses and conjunctive groups. Typically sentences are long and very complex, simple sentences are very rare. Punctuation plays a crucial role because legal texts usually include complicated syntactic patterns or long lists separated by semicolons. From the Natural Language Processing point of view, we consider a sentence to be the basic unit which is not true for texts that we have encountered in the environmental domain. Namely, we carried out information extraction with documents needed for the Environmental Impact Assessment (EIA) that considers the environmental impacts whether or not to proceed with a project. We designed and implemented the `EIA extractor` system to extract quantitative data from EIA reports, namely from the part on capacity of a given project. In general, its workflow is identical to the one of `RExtractor` but EIA reports are vastly different from legal documents, which requires to make some modifications in the architecture details: (1) EIA reports are structured using tables and lists and therefore we consider a noun phrase to be their basic unit and excluded a syntactic procedure from the processing in the NLP component, i.e., the reports are processed by a morphological tagging procedure only. (2) Thus we work with a linear representation of the sentence what drives us to use regular expressions for searching the EIA reports. The Database of Entity Queries was created manually by our partner in the INTLIB project and it contains several dictionaries, e.g., of entities, units, measures. The Database of Relation Queries consists of regular expressions for extraction of 4-tuples (*entity*, *quantity*, *amount*, *unit*), for illustration see the relations (*hall*, *area*, *96,000*, m^2), (*parking*, *capacity*, *1,150*, *slots*) in Example (13). We provide the details on the extraction in the methodology certified by the Ministry of the Environment of the Czech Republic (Borš et al., 2014).

(13) *Vlastní areál bude sestávat z halového objektu o ploše cca 96 000* m^2. *Předpokládají se 2 krytá stání pro jízdní kola a 1150 parkovacích stání.*

Lit. *The park will consist of a hall with the area of approx 96,000* m^2. *There will be 2 roofed bicycle parking stations and 1,150 parking slots.*

Contribution

We built the system that extracts domain-specific entities and relations from environmental documents. Their structure is very heterogeneous and we consider a noun phrase as a basic unit to process. Performance achieved by the `MST parser` on these documents is extremely low, so we processed them with the a tagging procedure only.

5. Future Perspectives

I entered the fields of corpus linguistics and natural language processing at the beginning of 1990s when Fred Jelinek and his team at IBM Thomas J. Watson Research Center invented a revolutionary method of statistical machine learning which was used for the first time in the task of machine translation (Brown et al., 1990). Since then I have participated in several major projects that have led to the development of this approach and thus to the progress in the field of Natural Language Processing (NLP). I contributed especially to (1) a substantial increase of the annotated data volume, (2) a formulation and implementation of alternative annotation strategy to decrease annotation costs, (3) a study of parsing procedure and its potential mainly for complex sentence parsing, (4) an exploration of annotated data to use them as a grammar workbook, (5) an exploration of understudied domains, e.g. the legal and environmental domains, and (6) a design of information extraction systems. Both the insight that I have gained so far and current trends have inspired me to identify directions of my next research:

(i) **NLP in Law** – In agreement with (Weischedel and Boschee, 2018) I still see a potential for building systems to extract knowledge bases from text and therefore I will try to contribute to further development of the `RExtrator` system. Definitely this development has to pay due respect to two revolutionary events in NLP: the framework of Universal Dependencies (UD) has joined the family of the corpus annotation frameworks and the methods of Deep Learning (DL) has grown up in the family of machine learning methodologies. These methodologies, both UD and DL have become a main framework/methodology of the NLP projects. In addition I want to initiate a discussion with legal experts to (i) review `RExtractor` and (ii) to assess the exploitation of parsing procedure for other legal documents than acts, e.g., contract cases (Durling, 2018).

(ii) **NLP in Digital Humanities** – The field of humanities has been enriched with the attribute 'digital' (DH) which is a consequence of large-scale digitization projects producing a plethora of texts. It opens up tremendous opportunity for NLP that should offer resources and tools for various requirements of humanities. I am particularly interested in which role NLP may play in Optical Character Recognition (OCR). Smith and Cordell (2018) state that "*There have been few efforts to think systematically or strategically about the problems of errorful OCR.*" Thus I have started to focus on improvement of statistical analysis of OCR output, see the pilot study of a student of mine (Nová, 2019).

(iii) **NLP in Education** – I see education as one of the application areas for advanced NLP techniques. My goal is twofold: (i) to turn the corpus-based exercise-book `Styx`, the `Čapek` editor, and the `PlayCoref` game into exciting resources for improving the learning of grammar and text comprehension experience; (ii) to explore the possibilities of the `RExtractor` system to be included in learning programs of law studies.

Bibliography

Eric Atwell, Geoffrey Leech, and Roger Garside. Analysis of the LOB Corpus: Progress and Prospects. *Corpus Linguistics*, pages 265–285, 1984.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0, 2013. URL `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Erik Borš, Radim Jäger, Tereza Jägerová, Martin Nečaský, and Barbora Hladká. Metodika pro automatizované inteligentní vytěžování nestrukturovaných dat v environmentální doméně, 2014. Methodology certified by the Ministry of Environment of the Czech Republic.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A Statistical Approach To Machine Translation. *Computational Linguistics*, 16(2):79–85, 1990. ISSN 0032-6585.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. Phrase Detectives: A Web-based Collaborative Annotation Game. In *Proceedings of I-SEMANTICS '08*, pages 42–49, Graz, Austria, 2008.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Jon Chamberlain, Christopher Cieri, and Karën Fort, editors. *Games4NLP: Using Games and Gamification for Natural Language Processing. A symposium co-located with the 15th EACL.* Valencia, Spain, 2017.

Jon Chamberlain, Udo Kruschwitz, Karën Fort, and Christopher Cieri, editors. *Proceedings of the LREC 2018 Workshop "Games and Gamification for Natural Language Processing (Games4NLP)".* Miyzaki, Japan, 2018. ISBN 979-10-95546-10-8.

Nancy A. Chinchor. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998. URL `https://www.aclweb.org/anthology/M98-1001`.

Fabio Ciravegna and Alberto Lavelli. Full Text Parsing using Cascades of Rules: an Information Extraction Perspective. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 102–109, 1999. URL `http://aclweb.org/anthology/E/E99/E99-1014.pdf`.

Robert Dale, H. L. Somers, and Hermann Moisl, editors. *Handbook of Natural Language Processing.* Marcel Dekker, Inc., New York, NY, USA, 2000. ISBN 0824790006.

James Durling. Diagramming Interpretation. *Yale Journal on Regulation*, (35):325 – 342, 2018. ISSN 0741-9457. URL `https://digitalcommons.law.yale.edu/yjreg/vol35/iss1/7`.

Jan Einarsson. Talbankens skriftsprakskonkordans, 1976a.

Jan Einarsson. Talbankens talsprakskonkordans, 1976b.

Stefano Federici, Simonetta Montemagni, and Vito Pirrelli. Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In *Proceedings of the Workshop on Robust Parsing*, 1996.

Erin Fitzgerald. Reconstructing Spontaneous Speech. Doctoral dissertation, Johns Hopkins University, Baltimore, Maryland, 2009.

Winthrop Nelson Francis and Henry Kucera. Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for use with digital computers, 1964. Providence: Department of Linguistics, Brown University.

Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. doi: 10.3115/992628.992709. URL `https://doi.org/10.3115/992628.992709`.

Jan Hajič, Petr Pajas, Pavel Ircing, Jan Romportl, Nino Peterek, Miroslav Spousta, Marie Mikulová, Martin Grůber, and Milan Legát. Prague DaTabase of Spoken Czech 1.0, 2017. URL `http://hdl.handle.net/11234/1-2375`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová Eva Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. Prague Dependency Treebank 3.5, 2018. URL `http://hdl.handle.net/11234/1-2621`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jan Hajič and Barbora Hladká. Probabilistic and Rule-Based Tagger of an Inflective Language: a Comparison. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 111–118, Washington, DC, USA, March 1997. Association for Computational Linguistics. doi: 10.3115/974557.974574. URL `https://www.aclweb.org/anthology/A97-1017`.

Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Hladká. *The Prague Dependency Treebank: A Three-Level Annotation Scenario*, chapter 7, pages 103–128. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003. ISBN 1-4020-1334-5.

Jirka Hana and Barbora Hladká. Getting more data – Schoolkids as annotators. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 4049–4054, İstanbul, Turkey, 2012. European Language Resources Association. ISBN 978-2-9517408-7-7.

Jirka Hana and Barbora Hladká. Universal Dependencies and Non-Native Czech. In Dag Haug, Stephan Oepen, Lilja Øvrelid, Marie Candito, and Jan Hajič, editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 105–114, Linköping, Sweden, 2018. Linköping University Electronic Press. ISBN 978-91-7685-137-1.

Jirka Hana and Barbora Hladká. CzeSL - Universal Dependencies Release 0.5, 2019. URL `http://hdl.handle.net/11234/1-2927`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jirka Hana, Barbora Hladká, and Ivana Lukšová. Sentence diagrams: their evaluation and combination. In Lori Levin and Manfred Stede, editors, *Proceedings of The 8th Linguistic Annotation Workshop (LAW-VIII)*, pages 38–47, Dublin, Ireland, 2014. Dublin City University (DCU). ISBN 978-1-941643-29-7.

Barbora Hladká and Jiří Hana. Parsing Writings of Non-Native Czech. In *Proceedings of the 4th Workshop on NLP Techniques for Educational Applications*, pages 12–16, Taipei, Taiwan, 2017. Asian Federation of Natural Language Processing. ISBN 978-1-948087-08-7.

Barbora Hladká and Ondřej Kučera. Prague Dependency Treebank as an Exercise Book of Czech. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Interactive Demonstrations*, pages 14–15, Vancouver, BC, Canada, 2005. Association for Computational Linguistics. ISBN 1-932432-55-8.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. Designing a Language Game for Collecting Coreference Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 52–55, Suntec, Singapore, 2009a. Association for Computational Linguistics. ISBN 978-1-932432-52-7.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. Play the Language: Play Coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 209–212, Suntec, Singapore, 2009b. Association for Computational Linguistics. ISBN 978-1-932432-61-9.

Barbora Hladká, Jiří Mírovský, and Jan Kohout. An Attractive Game with the Document: (im)possible? *The Prague Bulletin of Mathematical Linguistics*, (96):5–26, 2011. ISSN 0032-6585.

Barbora Hladká, Martin Holub, and Vincent Kríž. Feature Engineering in the NLI Shared Task 2013: Charles University Submission Report. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 232–241, Atlanta, Georgia, USA, 2013. Microsoft Research, Association for Computational Linguistics.

Barbora Hladká. Software Tools for Large Czech Corpora Annotation. Master thesis, Charles University, 1994.

Barbora Hladká and Jan Králík. Proměny Českého akademického korpusu. *Slovo a slovesnost*, 67(4):179–194, 2006. ISSN 0037-7031.

Barbora Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Votrubec. *Czech Academic Corpus 1.0 Guide*. Karolinum, Praha, Czechia, 2007. ISBN 978-80-246-1315-4.

Barbora Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. Czech Academic Corpus 2.0, 2008a. LDC - Linguistic Data Consortium, ISBN: 1-58563-491-3.

Barbora Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. The Czech Academic Corpus 2.0 Guide. *The Prague Bulletin of Mathematical Linguistics*, (89):41–96, 2008b. ISSN 0032-6585.

Barbora Hladká, Alevtina Bémová, and Zdeňka Urešová. Syntaktická proměna Českého akademického korpusu. *Slovo a slovesnost*, (4):268–287, 2011. ISSN 0037-7031.

Barbora Hladká, Ondřej Kučera, and Karolína Kuchyňová. STYX 1.0, 2017. URL `http://hdl.handle.net/11234/1-2391`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, copyright = Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0).

Irena Holubová, Tomáš Knap, Vincent Kríž, Martin Nečaský, and Barbora Hladká. INTLIB - an INTelligent LIBrary. In Karel Richta, Václav Snášel, and Jaroslav Pokorný, editors, *Proceedings of the Dateso 2014 Annual International Workshop on DAtabases, TExts, Specifications and Objects*, pages 13–24, Praha, Czechia, 2014. Czech Technical University in Prague, Faculty of Information Technology. ISBN 978-80-01-05482-6.

Pavel Ircing, Jan Švec, Zbyněk Zajíc, Barbora Hladká, and Martin Holub. Combining Textual and Speech Features in the NLI Task Using State-of-the-Art Machine Learning Techniques. In *The 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 198–209, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics. ISBN 978-1-945626-00-5.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009. ISBN 0131873210.

Marie Konárová. Školní větné rozbory jako možný zdroj závislostních korpusů (?). Master thesis, Charles University, Faculty of Mathematics and Physics, Czech Republic, 2012.

Vincent Kríž and Barbora Hladká. An Annotated Corpus Outside Its Original Context: A Corpus-Based Exercise Book. In *ACL 2008: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 36–43, Columbus, OH, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-08-4.

Vincent Kríž and Barbora Hladká. 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, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics. ISBN 978-1-941643-49-5.

Vincent Kríž and Barbora Hladká. Improving Dependency Parsing Using Sentence Clause Charts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop*, pages 86–92, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics. ISBN 978-1-945626-02-9.

Vincent Kríž and Barbora Hladká. Czech Legal Text Treebank 2.0, 2017. URL `http://hdl.handle.net/11234/1-2498`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vincent Kríž and Barbora Hladká. Czech Legal Text Treebank 2.0. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Vincent Kríž, Barbora Hladká, Martin Nečaský, and Tomáš Knap. Data Extraction Using NLP Techniques and Its Transformation to Linked Data. In *13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I*, volume 8856 of *Lecture Notes in Computer Science*, pages 113–124, Switzerland, 2014. Instituto Tecnológico de Tuxtla Gutiérrez, Springer International Publishing. ISBN 978-3-319-13646-2.

Vincent Kríž, Barbora Hladká, and Zdeňka Urešová. Czech Legal Text Treebank, 2015. URL `http://hdl.handle.net/11234/1-1516`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vincent Kríž, Barbora Hladká, and Zdeňka Urešová. Czech legal text treebank 1.0. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2387–2392, Paris, France, 2016. European Language Resources Association. ISBN 978-2-9517408-9-1.

Jan Králík and Ludmila Uhlířová. The Czech Academic Corpus (CAC), its history and presence. *Journal of Quantitative Linguistics*, (14, 2-3):265–285, 2007.

Karolína Kuchyňová. Porovnání větných rozborů, dokumentace k zápočtovému programu. Final project, Charles University, Faculty of Mathematics and Physics, 2016.

Ondřej Kučera. Pražský závislostní korpus jako cvičebnice jazyka českého. Master thesis, Charles University, Faculty of Mathematics and Physics, 2005.

Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. *Games with a Purpose (GWAPS)*. Focus series. ISTE Ltd, London, UK, 2015. URL `http://onlinelibrary.wiley.com/book/10.1002/9781119136309`.

Edith L. M. Law, Luis von Ahn, Roger B. Dannenberg, and Mike Crawford. Tagatune: A game for music and sound annotation. In *Proceedings of ISMIR*, pages 361–364. Austrian Computer Society, 2007.

Ivana Lukšová and Barbora Hladká. Information Extraction from EIA Documents, 2015. URL `http://hdl.handle.net/11234/1-1515`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19 (2):313–330, June 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972470.972475`.

Bohdan Maslowski. Automatické zpracování českých soudních rozhodnutí. Master thesis, Charles University, Faculty of Mathematics and Physics, 2015.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220641. URL `https://doi.org/10.3115/1220575.1220641`.

Raymond J. Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, June 2005. ISSN 1931-0145. doi: 10.1145/1089815.1089817. URL `http://doi.acm.org/10.1145/1089815.1089817`.

Olga Müllerová. *Mluvený text a jeho syntaktická výstavba.* Academia Praha, 1994.

Martin Nečaský, Tomáš Knap, Jakub Klímek, Irena Holubová, and Barbora Vidová Hladká. Linked Open Data for Legislative Domain - Ontology and Experimental Data. In *Business Information Systems Workshops*, volume 160 of *Lecture Notes in Business Information Processing*, pages 172–183, Berlin / Heidelberg, 2013. Uniwersytet Ekonomiczny w Poznaniu, Springer. ISBN 978-3-642-41686-6.

Jens Nilsson, Hall Johan, and Joakim Nivre. MAMBA meets TIGER: Reconstructing a Treebank from Antiquity. In *Proceedings of NODALIDA 2005 Special Session on Treebanks for Spoken and Discourse*, pages 119–132. Copenhagen Studies in Language 32, 2005.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phuong Lê H`ông, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Luong Nguy˜ên Thị, Huyen Nguyen Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. Universal dependencies 2.0, 2017. URL `http://hdl.handle.net/11234/1-1983`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kateřina Nová. A pilot study on evaluation of the Tesseract optical recognition engine. Final project, Charles University, Faculty of Mathematics and Physics, 2019.

Michal Novák. *Coreference from the Cross-lingual Perspective*. Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic, 2018.

Zdeněk Novák, Václav andŽabokrtský. Feature Engineering in Maximum Spanning Tree Dependency Parser. In *Lecture Notes in Computer Science, Vol. 4629, No. XVII, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, pages 92–98, Berlin / Heidelberg, 2007. Springer International Publishing.

Petr Pajas and Jan Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47, Genova, Italy, 2006. ELRA. ISBN 2-9517408-2-4.

Wenzhe Pei, Tao Ge, and Baobao Chang. An effective neural network model for graph-based dependency parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 313–322, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1031. URL `https://www.aclweb.org/anthology/P15-1031`.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, April 2013. ISSN 2160-6455. doi: 10.1145/2448116.2448119. URL `http://doi.acm.org/10.1145/2448116.2448119`.

Kiril Ribarov, Alevtina Bémová, and Barbora Hladká. When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion. *The Prague Bulletin of Mathematical Linguistics*, (86):21–38, 2006. ISSN 0032-6585.

Tjong Kim Erik F. Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 127–132, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: 10.3115/1117601.1117631. URL `https://doi.org/10.3115/1117601.1117631`.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht:Reidel Publishing Company and Prague:Academia, 1986.

David Smith and Ryan Cordell. A Research Agenda for Historical and Multilingual Optical Character Recognition. Technical report, Northeastern University, 2018. URL `http://hdl.handle.net/2047/D20297452`.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *CoRR*, abs/1805.11824, 2018. URL `http://arxiv.org/abs/1805.11824`.

Marie Těšitelová. *Kvantitativní charakteristiky současné češtiny.* Academia Praha, 1985.

Lucie Veselá. Systém STYX ve školní praxi: pilotní studie. Bachelor thesis, University of South Bohemia in České Budějovice, Czech Republic, 2012.

Luis von Ahn. Games with a Purpose. *Computer*, 39(6):92–94, June 2006. ISSN 0018-9162. doi: 10.1109/MC.2006.196. URL `https://doi.org/10.1109/MC.2006.196`.

Luis von Ahn and Laura Dabbish. Designing Games with a Purpose. *Communications of the ACM*, 51(8):58–67, August 2008. ISSN 0001-0782. doi: 10.1145/1378704.1378719. URL `http://doi.acm.org/10.1145/1378704.1378719`.

Luis von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems, volume 1 of Games*, pages 75–78. ACM Press, 2006.

Jan Václ. Tracing salience in documents. Master thesis, Charles University, Faculty of Mathematics and Physics, Czech Republic, 2015.

Ralph Weischedel and Elizabeth Boschee. What Can Be Accomplished with the State of the Art in Information Extraction? A Personal View. *Computational Linguistics*, 44(4):651–658, 2018.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anna Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. *Discourse and Coherence. From the Sentence Structure to Relations in Text.* Studies in Computational and Theoretical Linguistics. ÚFAL, Praha, Czechia, 2015. ISBN 978-80-904571-8-8.

Vladimír Šmilauer. *Novočeská skladba.* Kruh přátel českého jazyka, Praha, 1947.

A. Selected publications on Academic Corpus Annotation

B. Selected publications on Alternative Annotation

C. Selected publications on Information Extraction