HABILITATION THESIS

Barbora Vidová Hladká

Creating and Exploiting Annotated Corpora

Computer Science
Mathematical Linguistics

Prague 2019
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 Kucera, 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.*

![Prague Dependency Treebank annotation scheme](image)

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 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. I greatly appreciate having Jan Králik 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 Králik, 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

12004-2008, IET101120413, 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 Karolínka 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 (Kráž et al., 2015), (Kráž et al., 2016) contains the morphological and syntactic annotations that we en-

²2007-2008, 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. 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). 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

---

32012-2015, TA02010182, funded in 2012-2015 by the Technology Agency of the Czech Republic.

4http://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 to show Nicholson something, which he did and restored the actor’s 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 Kuchynková analyzed the diagrams that Marie collected (Kuchynková, 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 (UFAL). 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 Nečáský from KSI, the extracted knowledge was represented in the framework of Linked Open Data, see (Nečáský 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 parkovací 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 (Bors 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


DOI: –


DOI: [https://doi.org/10.3115/1631836.1631841](https://doi.org/10.3115/1631836.1631841)


DOI: –

Alternative annotation


DOI: [https://doi.org/10.3115/1698381.1698389](https://doi.org/10.3115/1698381.1698389)


DOI: [https://doi.org/10.3115/v1/N15-3005](https://doi.org/10.3115/v1/N15-3005)


DOI: [https://doi.org/10.3115/v1/W14-4905](https://doi.org/10.3115/v1/W14-4905)

Information extraction


DOI: [https://doi.org/10.3115/v1/N15-3005](https://doi.org/10.3115/v1/N15-3005)


DOI: [https://doi.org/10.18653/v1/P16-3013](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 Hajíčková. 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 Irčing from the Department of Cybernetics of the University of West Bohemia in Plzeň, Czech Republic and we took first place [Irčing 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 [Hajic 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.

<table>
<thead>
<tr>
<th>corpus</th>
<th>CAC</th>
<th>CLTT</th>
<th>STYX</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>original CAC</td>
<td>×</td>
<td>PDT</td>
</tr>
<tr>
<td>annotation scheme</td>
<td>PDT, UD</td>
<td>PDT, UD</td>
<td>sentence diagrams, UD</td>
</tr>
<tr>
<td>target annotation scheme</td>
<td>missing tokens, complex tokens, complex issues</td>
<td>spoken texts, complex sentences, nodes</td>
<td></td>
</tr>
<tr>
<td>exploitation of the corpus</td>
<td>tagging, cross-domain parsing, tagging,</td>
<td>parsing, information extraction, parsing</td>
<td></td>
</tr>
</tbody>
</table>

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 [Hajic and Hladka, 1997]. Its annotation was conducted in the De-
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 today's perspective it is useful to mention that the corpus annotation was a secondary goal of the quantitative analysis of Czech (Těšítelová, 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 Králik (2006) and Králik and Uhlírová (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 (Ríbarov 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).
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 consisting of 40,000 out of 1.3 million words.

(2) 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 té1, jako polonista, o tom1, no, ale nemohl to oddiskutovat, prostě říkal, že je to tedy2, ty otázky, tedy2 skutečně inflace, a ta hospodářská krize doléhá už právě nebo především také tedy2 na vědu a školství.

Lit. Professor Backvis was embarrassed by this, 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 the1, as a Polonist, about the1, well, but he could not discuss it, they just said, that this is thus2, those questions, thus2 indeed economic crisis already influences right now or above all also thus2 on science and education, thus2.

(3) I profesoru Backvisovi to bylo trapné, říkal, že jistě (tady ta) čeština by si zasloužila místo nejméně jako (ta) polština, (že, no tak,) že (mně to jako,) když mlouví on (o té1,) jako polonista (,) o tom1, (no,) ale nemohl to oddiskutovat, prostě říkal(i), že (je to tedy2,) ty otázky, tedy2 (skutečně) inflace(,) a (ta) hospodářská krize doléhá už právě nebo především také (tedy2) 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.

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), (Kriger 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). 

http://ufal.mff.cuni.cz/styx
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 REextractor system performance. REextractor 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 CLT 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). Afterwards CLTT was automatically morphologically and syntactically annotated using the MST parser (McDonald et al., 2005) adapted for Czech by Nováček (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.

**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. 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 Stepánek, 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:

- 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

---

8English 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) \( S \) výjimkou ukončení činnosti mohou účetní jednotky podle §1 odst. 2 písm. d) až h) ukončit vedení účetníctví nejdříve po uplynutí 5 po sobě jdoucích účetních období, ve kterých vedly účetníctví.

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.

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.

<table>
<thead>
<tr>
<th>annotation</th>
<th>academic</th>
<th>alternative</th>
<th>alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>CAC, CLTT, STYX, PlayCoref, Čapek</td>
<td></td>
<td></td>
</tr>
<tr>
<td>format</td>
<td>research project</td>
<td>game</td>
<td>language class</td>
</tr>
<tr>
<td>team</td>
<td>linguists</td>
<td>players</td>
<td>students</td>
</tr>
<tr>
<td>required knowledge</td>
<td>annotation</td>
<td>language</td>
<td>language</td>
</tr>
<tr>
<td>instructions</td>
<td>framework</td>
<td>knowledge</td>
<td>grammar</td>
</tr>
<tr>
<td>environment</td>
<td>editor</td>
<td>game</td>
<td>editor</td>
</tr>
<tr>
<td>gold annotation</td>
<td>arbiter</td>
<td>agreement</td>
<td>combination</td>
</tr>
</tbody>
</table>

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...
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]. 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. 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.

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.

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 stars Jack Nicholson as Randle McMurphy. As Forman did not allow the actors to see the day’s filming, this led to the cast losing confidence in him, while Nicholson also began to wonder about his performance. Douglas convinced Forman to show Nicholson something, which he did, and restored the actor’s confidence. (Source: Wikipedia)
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). 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 Dependency Treebank is the only corpus where coreference is annotated on the underlying dependency-based syntactic layer (t-layer), see e.g., (Zíkanová 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 (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 $\lambda_1 \times F(A, \text{acr or gold}) + \lambda_2 \times F(A, B) + \lambda_3 \times \min(12, s)/12$, 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.

http://www.phrasedetectives.org
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 Hladka, 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 (Krůž 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
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)
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) Účetní jednotky, které nejsou organizační složkou státu, územním samosprávním celkem nebo účetní jednotkou vzniklou nebo zřízenou zvláštním zákonem, mohou uplatnit hospodařský 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.
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-
Table 4: Unlabeled Attachment Score of the MST parser and the Clause Chart Parsing procedure (CPP).

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

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), shallow 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 \(0, M\), 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.
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 OB0 and OB1 patterns, e.g., OB0, OB1, OB1B0, OB0B0. 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 (Kříž 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 REextractor 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, m²), (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 m². 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 m². 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 **RExtractor** 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


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.


Karolína Kuchynová. Porovnání vět ních rozborů, dokumentace k zápočtovému programu. Final project, Charles University, Faculty of Mathematics and Physics, 2016.


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.


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.


A. Selected publications on Academic Corpus Annotation
B. Selected publications on Alternative Annotation
C. Selected publications on Information Extraction