INSTITUT FÜR
COMPUTERLINGUISTIK

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Universität Heidelberg, Institut für Computerlinguistik, INF 325, 69120 Heidelberg

Heidelberg, 12.01.2020

Prof. Dr. Anette Frank
Computational Linguistics
Tel. +49 6221 54-3247
Fax +49 6221 54-3242
frank@cl.uni-heidelberg.de

Review of the Habilitation Thesis entitled "Creating and Exploiting Annotated Corpora" offered by Barbora Vidová Hladká in 2019

The basis for my review is the text of the habilitation thesis of Dr. Barbora Vidová Hladká *"Creating and Exploiting Annotated Corpora",* dated 2019. As the title clearly reflects, the thesis centers around annotated corpora and covers research questions regarding their creation and their exploitation for various applications. The thesis structures these research questions into three areas: (A) **Academic Corpus Creation**, (B) **Alternative Annotation**, and (C) **Information Extraction** as one of many possible use cases that can be built on the basis of annotated corpora.

Structure of the thesis

For each of the three research areas Dr. Hladká selected and included 2-3 of her own publications that she considers most representative and important for her research on these respective themes.

The full thesis starts with an **overview** that introduces the individual themes and how they are linked to each other, and in which the author situates her own work and contributions within the research field. This overview situates the work covered in **the eight enclosed publications** and ca. **12 additional publications** of the author in these respective areas. The included eight publications are mainly co-authored. For each of them the contribution of Dr. Hladká is quantified. Her contribution ranges from 40-70%, in most of the cases showing significant contribution (65-70%). The venues involve *ACL workshops that are specialized towards corpus annotation, methods and applications for computational language learning, LREC publications, or *ACL demo publications to publicize and showcase practical annotation and NLP tools.

We also learn that beyond the work covered in the thesis, additional research of Dr. Hladká targets the interesting task of Native Language Identification from expression in a second language for written and spoken language, for which she achieved a first place in a community shared task.

Following the overview, for each area **a detailed summary** is given of the research questions and contribution of her own work in the various themes, on 12, 4 and 8 pages, respectively.

Universität Heidelberg
Institut für Computerlinguistik

Im Neuenheimer Feld 325
69120 Heidelberg
www.cl.uni-heidelberg.de/~frank

Summary of research contributions

The thesis of Dr. Hladká covers important contributions to **(A) research on and the practical creation of linguistically annotated corpora** with the aim of supporting theoretical and empirical studies, and as a basis for building Natural Language Processing tools. She contributed to the linguistic (re)annotation and conversion of important corpora of the Czech language, including one of the earliest annotated corpora, the CAC Corpus (Czech Academic Corpus) and of the CLTT (Czech Legal Text Treebank). The CAC covers levels of linguistic annotations that range from morphological over syntactic and semantic structures. The work of Dr. Hladká mainly focuses on morphological and syntactic annotation. It covers annotation description, conversion to new annotation formats, the development of annotation tools and training of automatic labeling systems. She explored the exploitation of these corpora to construct a corpus-based *Exercise Book of Czech*, the STYX, to support language learning. She further contributed to the creation of the CLTT corpus that covers text from the legal domain and that includes morphological, syntactic as well as entity and semantic relation information.

Since manual linguistic annotation is extensive and time-consuming, research aims to **(B) harvest linguistic annotations in crowd-sourcing settings**. Dr. Hladká investigated *Games with a purpose* (GWAP), in which target annotations are derived as a secondary product in specifically designed interactive online games. She showed how to port the GWAP setting from the annotation of images to the annotation of texts, by developing the PlayCoref game to collect coreference annotations on text. She further developed a tool to enable students to perform morphological and syntactic annotations by creating sentence diagrams with an editor designed for this purpose.

Finally, her research covers **(C) the exploitation of linguistically annotated corpora for use in NLP applications**. She concentrates on Information Extraction, a classical NLP application task that exploits linguistic annotations in corpora and learns to automatically annotate entities and relations in natural language texts, and to store this structured information in a knowledge base. She worked on two challenging language genres: legal and environmental texts. She developed entity and relation extraction tools to perform **information extraction** on legal documents and to **construct knowledge bases** covering relations such as definition, right and obligation (Kriz and Hladká, 2015), addressing the challenging problem of domain adaptation (i.e., adapting the tools trained on PDT to the much longer sentences and complex constructions of the legal language documents). To enhance the quality of the syntactic parses underlying the IE and KB construction tasks, she studied strategies to overcome the challenge of long-winded, complex sentences by a split-parse-and-recombine approach (Kriz and Hladká, 2016). Further work includes the extraction of quantitative data from environmental documents using tools that apply regular expressions to morphological annotations.

The overview of the work conducted in the three research areas is clearly structured and shows how the work of Dr. Hladká is embedded in the larger scientific context. It shows an impressive coverage and a wide variety of ways in which she contributed to the creation and continuous development of annotated resources for the Czech language. These contributions range from deep investigation, harmonization and re-transformation of annotations for sustainable usage to the development of enhanced annotation methods using crowdsourcing and applications for computational language learning and NLP applications that make use of the morphological and syntactic annotations to extract information from documents for the automatic creation of knowledge bases. Both the early work of Dr. Hladká on automatic language learning tools and more recently the application of Information Extraction to legal and environmental domains is in line with the growing impact of such applications in our field.

Detailed discussion of the work in areas A to C and the contributions of Dr. Hladká is provided in the subsequent individual summaries.

(A) Academic Corpus Annotation covers the academic annotation of the three Czech corpora: CAC, CLTT and STYX. The much older CAC corpus was merged with the Prague Dependency Treebank (PDT) to increase its coverage. To do so, its annotations had to be transformed or completed to be harmonic with the PDT annotation style. Upon this transformation, the CAC and PDT annotations

were further transformed to the UD annotation scheme that has now prevalent. The integration of CAC in PDT was deeply studied in **Hladká et al. 2011**, to identify ways of integrating the older CAC annotations in PDT. Since written and spoken language differ in many ways, and spoken language phenomena were only partially annotated in CAC, only the annotations over the written language data was transformed with a mix of automatic conversion and manual additions.

Kriz and Hladká 2008 use the annotated PDT to create a tool for grammar teaching of Czech, by extracting sentences from PDT and transforming them to sentence diagrams as used in teaching, thus creating a *Corpus-based Exercise Book of Czech,* called STYX. An editing tool allows for selection of sentences that satisfy specific syntactic properties, the interactive construction of sentence diagrams and automatic correction against gold annotations. Transformation rules have been defined over morphological and syntactic PDT annotations that convert them to appropriate sentence diagrams.

To support NLP applications such as IE in legal domains, **Kriz and Hladká 2018** create the CLTT (Czech Legal Text Treebank) over legal documents. They perform morphological and syntactic annotations according to the PDT scheme and further enrich them with entity and relation annotations. The corpus is used to train an IE system (RExtractor) that applies to dependency trees in order to extract entities and relations. Both manual and automatic syntactic annotation of legal documents is challenging given the high complexity of structure and tokens and the extreme length of sentences in legal documents. Kriz and Hladká design rules that split sentences and join tokens into larger units, and run automatic pre-annotations with the MST parser trained on PDT on the smaller subsequences. Annotators check and correct the individual syntactic trees and add inter-segment links that connect the validated subtrees into complete structures. Annotated entities and relations are projected on the dependency structures in the form of Subject-Predicate-Object relations and can be searched with tree-queries. The produced annotations in CLTT are later used for evaluation in the IE task (C). Overall, the CLTT contributes to the scale and variety of morphologically and syntactically annotated data for Czech, and supports downstream NLP tasks in novel domains.

(B) **Alternative annotations** are considered in tasks that require annotations at low costs. For this, crowdsourcing has been studied, as well as games with a purpose GWAP. The specific challenges here are to make the tasks accessible for untrained annotators, often by breaking the task down into simple steps, to control of the quality of the produced annotations and to recruit annotators in case they are not paid. **Hladká et al. (2009a,b)** apply GWAP to coreference resolution. They perform some pre-annotation, e.g. to block tokens that are not a valid annotation target using morphological information. They also include reference annotations from annotated data to be able to rate the reliability of annotators. For this, they can draw on annotated instances from PDT. The reported work represents one of the earliest attempts to apply GWAP on textual data. It is at the level of a pilot study that does not include final data evaluation results. One of the problems here is to attract players for a game on texts, which is not easy to achieve.

Thus, in another setting, the target groups for crowdsourcing are school classes, where one can, e.g.., build on the students' grammatical knowledge. **Hana et al. 2014** develop a crowdsourcing tool for sentence diagramming for use in school classes. This scenario has a dual usage: it can be used for training grammar analysis skills in school and for harvesting annotations that the students produce. Hana et al. develop an online editor for sentence diagramming that is the reverse of the STYX tool discussed above and implement an evaluation metric that compares annotations using tree edit distance. They also develop an algorithm for merging trees. For evaluating annotations, on can rely on multiple annotations by students and use e.g. majority voting, and reference annotations provided by teachers. In pilot experiments Hana et al. measure overlap of annotations between pairs of teachers, pairs of students and pairings of teachers and students and find high overlap between teachers and little overlap between students. Thus, teacher annotations seem to be instrumental. The scenario is interesting, also in view of automatic grading in Computer Aided Language Learning (CALL). However, one needs to take into account the number and the complexity of annotations one can expect to collect in training or grading scenarios, and hence a more promising application of the presented research is in my view to be seen in the context of automatic grading in CALL.

In part **(C),** the thesis covers the exploitation of annotated corpora for downstream NLP applications, in particular, **Information Extraction.** This task is explored in legal and environmental domains. Specifically, **Kriz and Hladká 2015** develop the RExtractor system with the aim of studying the benefits of performing entity and relation extraction from textual documents based on morphological and syntactic annotations. The system uses automatic morphological and syntactic parsing tools, MorphoDita and the MST parser trained on the PDT. Building on the entity and relation annotations performed on the CLTT corpus (see A), these entities and relations are stored in databases and provide the targets for detecting them in the automatically parsed documents. The annotated CLTT then serves as evaluation corpus for the IE system. The authors design the RExtractor system that applies manually defined extraction rules to the parsed texts, using a tree query language. The authors study the effect of such rules on the complex language encountered especially in legal texts and showcase the advantages of using syntactic annotations, especially when it comes to coordinated structures or deeply embedded constructions. The enclosed publication presents the design and technical details of the system, it illustrates how extraction is performed by defining and running extraction rules, and illustrates its advantages in application to selected cases.

As already mentioned by Kriz and Hladká 2015, the performance of a syntactically grounded information extraction system crucially relies on the quality of automatically constructed parses. Given the complexity of legal texts especially in view of the overall sentence length and complexity of syntactic structure, the application of such a system to legal texts presents a true challenge. Hence, **Kriz and Hladká 2016** explore the potential of a split-parse-and-join approach for parsing that is intended to first parse subsequences of smaller segments of the sentence, and to integrate these structures into the larger sentence structure in a second step. The intuition being that parsing on smaller sequences yields better results, and that systems can profit from such factorization for producing overall better results. Kriz and Hladká 2016 perform experimental studies to confirm this hypothesis on data from the PDT and CAC sections of the general domain corpora. They propose the so-called CCP (clause chart parsing) approach that performs parsing in a two-stage process: first coordinated and subordinated clauses are parsed separately with respect to the sentence clause chart and only in the second step their dependency trees become subtrees of the final sentence tree. Kriz and Hladká establish experimentally on PDT data that CCP obtains an improvement of UAS of ca. 1 percentage point, by factorizing coordinated and subordinated structures. Thus, the factorization of sentences into subclause units in a chart-based approach was shown to be beneficial for overall parse quality and especially for complex sentences. Performing factorization within the chart avoids a possibly erroneous pre-processing step for sentence splitting. In addition, the chart-based approach is very general and can be applied to different languages and parsing algorithms.

In summary, the thesis of Dr. Hladká covers the **complete life-cycle of corpus annotations** that starts from the process of performing manual annotations on text corpora over the design of alternative annotation methods to the exploitation of the annotations in downstream tasks. This comprises important contributions to the **study, creation and integration of manually performed annotations**, including making existing annotations sustainable by porting them to newer annotation schemes, and the creation of novel and extended annotations on texts of novel, challenging and understudied domains, such as legal texts that are starting to gain importance in the field.

Dr. Hladká shows the **potential of academic corpus annotation to applications in language learning** by creating STYX, a corpus-based extraction of grammatical sentence diagrams to be used in teaching contexts. She further **pioneered the investigation of novel forms of annotation using crowdsourcing**, targeting two phenomena related to academic corpus annotation: coreference resolution and sentence diagram creation in classrooms to harvest annotations. While harvesting of novel and reliable annotations remains difficult in both scenarios, the latter application is a promising technique in the context of CALL. The contributions in this area are clearly novel insights, as well as the creation of editing tools that can be used or extended in other contexts.

Finally, Dr. Hladká contributed to the **study and creation of tools for information extraction**, by showing the potential benefit of employing automatic syntactic annotation as a basis for information extraction with the RExtractor system. She further addresses challenging problems in the automatic

parsing of complex language in the legal domain, by proposing a novel chart-based parsing strategy that factorizes the parsing process into sub-sentential components and which was experimentally shown to be effective in studies on the PDT.

Overall, the thesis of Dr. Hladká is very clearly structured and presented. The thesis forms a coherent research theme that integrates several contributions that together support the complete life-cycle of linguistic annotations: this cycle starts with the process of manual corpus construction and annotation design (A), which can be extended to alternative, less cost-intensive annotation settings (B) that, however, must ensure backwards compatibility of new annotation instances obtained in subsidiary tasks with the original annotation schemes. Likewise, the thesis shows how the original annotation design can be straightforwardly exploited in downstream tasks (C), such as Information Extraction, where annotations that extend the original annotations can be used for training and evaluation of downstream NLP tasks, and where we need additional annotation coverage on novel language genres to help adapt the pre-trained parsing models to support NLP tasks in novel application domains.

The thesis comprises research results of eight publications that have been peer-reviewed in specialized workshops in renowned international conferences. The obtained research results deliver new insights in methods for corpus construction and their maintenance, their extension to application tasks such as computer-aided language learning (CALL), as well as methods for alternative acquisition of annotations through crowdsourcing or CALL applications; they comprise novel language resources and various tools that are instrumental for extended application contexts, such as Information Extraction and CALL that build on the original annotation design. Finally, Dr. Hladká contributes to the study of domain adaptation of parsing systems, again taking advantage of syntactic annotations in annotated treebanks. The contributions of Dr. Hladká cover a wide range of techniques in a strongly intertwined research area. They are clearly important for the advances of NLP for the Czech language, but are also of wider interest, since they are extensible to other languages or application contexts.

I strongly recommend the acceptance of the thesis of Dr. Hladká and the attribution of the corresponding title of an associate professor (or docent).

With best regards,

Prof. Dr. Anette Frank