Morphologically and Syntactically Annotated Corpora of Many Languages

Current State, the Problem and its Significance

Annotated corpora have become a standard resource for research in both linguistics and computational processing of natural languages. Lexicographers judge word usage and distribution by occurrences in corpora; part-of-speech tags may help them narrow their queries. Grammarians may use syntactically annotated corpora *(treebanks)* for queries such as "show me all examples where a verb governs two objects in the accusative." In natural language processing (NLP), syntactic parsing is an important preparatory step for many tasks such as question answering, data mining or machine translation; the state-of-the-art parsers rely on human-annotated treebanks and apply machine learning algorithms to extract linguistic knowledge from the treebanks.

Despite the costs and qualified labor needed to create them, treebanks are available for many languages these days. They range from small proof-of-concept sets of just a few hundred sentences, to monumental works with tens of thousands of sentences and millions of words (such as the Penn TreeBank or the Prague Dependency Treebank[1]). Unfortunately, size is not the only parameter that varies. The corpora were created by many different institutes and teams around the planet. Some teams adapted annotation guidelines from other projects, other teams developed their own and unique. As a result, language phenomena that are similar or equivalent in two languages, are often captured in very different, even incompatible ways. Two treebanks for the same language are sometimes incompatible too, if they were created by different teams with different objectives and standpoints.

To give just a few examples:

- Pronouns, determiners, pronominal numerals and pronominal adverbs form a vague group of many faces. Every corpus seems to use unique definitions and rules to classify words in this group and assign part-of-speech tags to them.
- Numerals constitute a separate part of speech in some treebanks while they are considered adjectives in others (ordinal numbers in English Penn Treebank; all numerals in Danish Dependency Treebank).
- Many treebanks show nouns within a prepositional (or postpositional) phrase as dependents (child nodes) of the preposition. If there is also an adjective and/or determiner (as in "on the new bridge"), they are attached to the noun. However, in the German Tiger treebank, all three words are child nodes of the preposition. In the treebank of Hindi, the adjective will depend on the noun but the noun will also govern the postposition.
- There are at least four main approaches, and numerous variants, how to capture coordinate structures within a dependency tree.

Obviously this sort of diversity has unfortunate consequences.

Cross-language studies are difficult, if not impossible. Suppose we want to see what language is more prone to using non-projective constructions (see Chapter 12 in Zeman 2004). How can we know, if the annotation guidelines used in one of the treebanks systematically lead to technical non-projectivities (such as in attaching punctuation)? Suppose we have ten treebanks of ten different languages and want to compare them based on complexity of embedded relative clauses. How long will it take to study the tagsets of all involved corpora in order to recognize a relative clause, or even a verb? Not to mention that documentation of some of the corpora is not easy to find, or it is available in the local language only.

Published results of natural language processing tools, such as syntactic parsers, are not comparable. Not just across languages—they may be incomparable between two corpora of the same language! If one of the corpora chooses to capture a phenomenon in a way that is more difficult for the parser to learn, the accuracy of the parser will invariably decrease. Sometimes it may be just because more information is captured. In many cases however, there are two or more ways of expressing the same syntactic relation that are equivalent

1 We do not provide references for every corpus mentioned in this proposal. Instead, please refer to Zeman et al. (2012) where all the corpora are described and proper references provided.

in expressive power (meaning that we can transform the representations without loss of information) but they are not equivalently easy from the perspective of a machine-learning algorithm.

There have been a few attempts to address the issue from various perspectives. Tsarfaty et al. (2011) proposed a parser evaluation technique that was robust with respect to some (but not all) annotation variations. Bosco et al. (2010) describe and evaluate (by parsing) divergences between two Italian treebanks. Other authors (Nilsson et al. 2006, Bengoetxea and Gojenola 2009) experimented with transformations of structural annotation (i.e. labeled parent-child links between tree nodes) with the aim of improving parsing accuracy. This work was limited to one or two languages and a very narrow selection of phenomena. McDonald et al. (2011) proposed a *Universal Part-of-Speech Tagset*; they used it for cross-language parser training. They did not transform structure however, and they observed that different annotation schemes across treebanks were responsible for the fact that some language pairs worked better together than others. In a recent follow-up they propose a similar approach to syntactic annotation (McDonald et al., 2013). Schwartz et al. (2012) defined two measures of syntactic *learnability* and tested them with five different parsers on varying annotation styles of six phenomena (coordination, infinitives, noun phrases, noun sequences, prepositional phrases and verb groups). They worked only with English and they generated varying annotations during conversion of constituency annotation of the Penn Treebank WSJ corpus to dependencies.

A common feature of all the contributions mentioned so far is that they focus on syntactic parsing, i.e. on the problems that varying annotation causes to the area of NLP. They limit themselves to just one or a few languages (with the exception of the Universal POS Tagset) and to just a few phenomena. Benefit for corpus-based linguistic research is negligible.

We believe that all the problems, parsing and linguistic alike, will only be eliminated if we are able to convert every treebank to one common annotation style. Better yet if this is not necessarily one fixed style but if we provide means for the researcher to pick a style that they favor over the others. As long as all data in all treebanks can be brought under the same guidelines, it is OK.

The proposer of this project has already been active in this area. So to conclude the survey of related work, we should also cite *Interset* (Zeman 2008). Interset is a universal set of morpho-syntactic features (part of speech, gender, number, tense etc.) and their predefined values. It is meant as a sort of Interlingua for morphological tagsets. One could in theory take any part-of-speech or morphological tagset of any language and define procedure that would convert the tags to Interset. (Obviously it is possible that the need for new Interset features and/or values will be discovered occasionally and they will be added.) Together with a sophisticated value-replacement algorithm, conversion procedure between any two Interset-mapped tagsets can be derived, although some features will probably get lost during the conversion *from Interset* to the target tagset (depending on the set of features that the target tagset can accommodate). Despite the names, Interset is quite the contrary of the Universal POS Tagset. McDonald et al. did not need to study the details of existing tagsets too carefully because they drop most of the information, except for the core parts of speech that they find universal enough. In contrast, Interset drops as little as possible ("universal" in Interset means "keep what you find anywhere", not just what you find "everywhere").

In the area of structural normalization, we have proposed a Harmonized Multi-Language Dependency Treebank—*HamleDT* (Zeman et al. 2012). Conceptually, HamleDT is a direct predecessor of what we are proposing to do in this project. Its potential beneficial effect was confirmed by the warm response we got at the LREC 2012 conference where the idea was presented. Practically however, there is still a long way from HamleDT in its current state to what we believe would bring a real breakthrough to the methodology of various types of crosslingual research in computational linguistics.

The trouble with both Interset and HamleDT is that they emerged as by-products of projects whose main goals were something else: parser adaptation in the first case, and machine translation in the latter. We had to deal with varying annotation styles of multilingual data and we gave a thought to how such interoperability issues should be addressed systematically. The prototype of HamleDT proved viability of the idea and taught us a lesson; nevertheless, there are still more questions than answers. For many languages in HamleDT, morphological conversion to Interset has been just sketched; the same holds for studying syntactic tags (labels of dependencies); a few varying structures have been identified but their normalization has not been defined for all languages; and some complex phenomena, such as verb groups, have not yet been touched anywhere. That is why we believe that there should be a research project devoted to this idea. What we propose is to study a vast majority of currently available dependency treebanks, provide a comprehensive

description of their annotation styles and reshape the idea of HamleDT into a highly useful resource for both NLP and corpus-based studies.

General demand for such sort of data can be indirectly illustrated by the number of citations of papers that describe them—for the multilingual set from CoNLL 2009 Shared Task it is (according to Google Scholar) 190 citations, for PDT about 500, other treebanks co-created at the proposer's institute have roughly 200. PDT alone has 383 registered users worldwide (before version 2.5, which does not require registration any more).

The results will be useful to some extent even for treebanks that will emerge after this project is finished. Methods, transformation procedures and infrastructure will be ready. Chances are that a significant portion of annotation styles of new treebanks will match something that we will have dealt with previously. The tedious work of studying documentation and the data cannot be skipped but the rest will be easier than now, and the new corpus will instantly become compatible with 30 others, not just with one selected.

We are not aware of any other previous research that would have brought together so much data for so many languages and compared them in such a thorough manner.[2] There is a chance that the research will shed some light on universal patterns and principles that remain hidden when one studies just one or a few corpora. This affects both the annotation per se, and the problem of finding the most effective and accurate way how to design the transformations.

To avoid misunderstanding, let us also stress what we are **not** going **to do.** We are not going to propose any international standard for annotation styles or data formats. There have been attempts to standardize linguistic annotation (TEI Consortium, EAGLES 1996) and the International Organization for Standardization slowly works on others (Windhouver 2012). Some existing corpora meet the standards, which somehow facilitates using them, others do not. Rather than designing or improving standards, we want to find a good way of using the data that just "is out there"—regardless what standard they meet, if any.

We are also not going to research *the languages*. We are going to research *existing annotated data* for the languages. Providing a full description of morphology and syntax in 30 languages (some of which we do not understand) would indeed be a task beyond the potential of a tiny research team within a three-year project.

Project Outline and Expected Outcomes

Normalization: We will explore morphologically and syntactically annotated corpora (treebanks) of 30 (or more) languages belonging to several language families. We will make an inventory of phenomena annotated in the treebanks, and we will compare the means of representing the phenomena *(annotation style* or *scheme)*. We will assess mutual convertibility between the various means found to capture the same phenomenon. Then we will propose a universal annotation style to which all the original annotation styles could be converted with minimal loss of information; naturally, we will also propose, implement and evaluate the conversion procedures.

Transformation: We will identify syntactic structures that appear in many treebanks and their annotation style differs across treebanks. We will design transformation procedures that switch between annotation styles of these *varying structures*. We will evaluate them with multiple statistical parsing algorithms and assess *learnability* (Schwartz et al. 2012) of the alternative styles.

Publication: Selected research results will be published during all three years of the project as conference papers and/or journal articles. We expect a minimum of two papers/articles per year (see also part C2). At the end of the project, the findings will be summarized in a monograph (in English) that will provide a comprehensive overview of annotation styles currently used in treebanks around the world. Such a book will serve as a reference for linguists-users of corpora. At the same time it shall provide an overview of existing issues and their solutions, useful for linguists-creators of future corpora.

We plan on presenting our results at important international meetings such as ACL or COLING; the biennial conference on Language Resources and Evaluation (LREC) is also highly relevant to the proposed research and we hope to get a paper accepted in 2016 (the only LREC year falling into the duration of the project).

2 One large multilingual data collection we are aware of is the InterCorp corpus (see References). It is normally available only through a web service (no direct download of the data) and its annotation does not go beyond morphological tags.

Text, Speech and Dialogue (TSD) is our likely target among domestic conference series. We will also submit an article to Computational Linguistics (J_{imp}).

We will make the normalized annotations available to the research community in a way allowed under the license terms of the original treebanks. Transformation procedures will be made available as well, so that researchers can mix their own unified annotation style if the normalization proposed by us does not suit their needs. As we do not own the IPR (intellectual property rights) for the original treebanks, the way of making the results available will vary. 13 treebanks (out of the 30 we plan to process) come with free non-commercial licenses under which we can (and will) redistribute the original data together with our modified annotation. 11 treebanks are easily obtainable for free but the users have to obtain them directly from the original sources; we will provide "patches" that the users can use to get the normalized annotation once they have the data. The remaining 6 treebanks are either available for a fee or they lack a regular distribution channel; we will provide the patches for the users who have them, and trained parsing models for the others. Needless to say, whatever we add to the data (and thus hold copyright for) will be provided free-of-charge for non-commercial research, namely under the CC-BY-NC-SA license. The data we create will be stored in and distributed through the LINDAT/CLARIN repository (**http://lindat.mff.cuni.cz/**) and assigned a persistent URL (**http://handle.net/**), which guarantees permanent availability of the resources after the project termination, regardless of the presence of the team members at the institute.

Methods and Strategy

We are mainly interested in dependency syntax. However, we do not a priori exclude constituency treebanks. Dependencies and constituents are two interconnected faces of one system. Where possible, we will use head-selection tables to convert constituency treebanks to dependencies (Johansson and Nugues 2007). Then the normalization will for the most part be already defined by the conversion process. (Note that some other treebanks that we intend to work with were converted by other people from constituents to dependencies.)

The 30 treebanks we plan to work on (and have access to) comprise the following languages: Arabic, Basque, Bengali, Bulgarian, Catalan, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, German, Greek (Ancient), Greek (Modern), Hindi, Hungarian, Italian, Japanese, Latin, Persian, Portuguese, Romanian, Russian, Slovenian, Spanish, Swedish, Tamil, Telugu, Turkish. We deliberately limit this list to a manageable size but we know about other treebanks that we may add if time permits (for example Hebrew, French, Polish, Slovak; 11 languages currently covered by McDonald et al. (2013); 11 tiny treebanks of extinct languages in the PROIEL project etc.) Different corpora of one language are also an option, e.g. for Chinese.

A typical dependency treebank contains the following human-assigned values for every node (which corresponds to a word or other token):

- the word form
- lemma (not available for all treebanks)
- morphological tag (part of speech tag + values of morphosyntactic features)
- link to the parent node
- syntactic tag, i.e. label of the dependency relation between the node and its parent

In addition, some treebanks contain NULL nodes that correspond to elided words, i.e. they do not correspond to any token physically present in the annotated sentence. And some treebanks collapse selected multi-word expressions into one token.

Thus there are three broad areas for possible normalization:

- *morphology* … to this level we count converting the *morphological tag* (which is a combination of part of speech tag and morphosyntactic features) to a universal representation (Interset)
- *syntax* … any structural changes, i.e. changing parent-child links and/or their labels *(syntactic tags)* belong here
- *tokenization* … we could also experiment with the differing approaches to NULL nodes and multi-word expressions

We plan on using and extending the idea of Interset (Zeman 2008) on the morphological level. The cornerstone of Interset is a universal set of features capable of capturing all information that authors of corpora store in morphological tags. There is a set of possible values for every feature (e.g. "sing" (singular) is one of the possible values of the feature "number"). The sets of features and values are open because we may encounter new features as we cover new languages. Nevertheless, we should always ask first whether the new feature is not something that can already be represented: sometimes it has just a different name in the grammar of the new language. Interset also has means of converting a feature structure to another "physical" tagset. Such process is by definition lossy and there are sophisticated algorithms that select features fitting the target tagset. This part of Interset is less important, as we primarily want to translate annotation from a physical tagset to Interset, which is our normalized form. However, it may be useful as a first approximation for a linguist who is familiar with one tagset and who wants to explore data tagged with different tagset. For illustration of Interset, consider two tagsets of Czech, one used in the Prague Dependency Treebank, the other used by the morphological analyzer Ajka (Sedláček and Smrž 2001). The following are tags assigned to the Czech word *seděl* ("he sat"). Underlined are Interset features unique to one of the sources.

Ajka	**PDT**
`k5eAaImAgMnS`	`VpYS---XR-AA---`
Interset	**Interset**

`pos = verb, verbform = part,`
`aspect = imp, tense = past,`
`gender = masc, animateness = anim,`
`number = sing,`
`negativeness = pos`
	`pos = verb, verbform = part,`
	`tense = past, voice = act,`
	`gender = masc, number = sing,`
	`negativeness = pos`

The set of Interset feature-value pairs is the normalized target form we want to arrive at. Note that in the example above, we still will not have two identical representations, because of the unique features on either side. Ideally, we would want to get the *union* of the two feature-value sets. But we can create it only if we get the two different annotations for the same corpus—which is typically not the case. If we have two different corpora of one language, we might want to automatically extract a wordlist for the language, extract Interset annotations, apply a statistical disambiguation model (possibly error-prone) and create the union. However, we will be prevailingly dealing with data in different languages, which makes such unification impossible.

Applying Interset to tagsets of 30 languages involves different types of work. Most of the labor and time will go to decoding the current tags of the physical tagset, writing filters that convert them to the existing features of Interset, and testing the filters. Occasionally we may encounter a feature in the physical tagset for which we will have to decide whether it is representable by the current Interset features, or we have to define a new feature in Interset. However, there are also important conceptual questions that have to be solved in Interset (they should have been solved by design but unfortunately we have not foreseen them until we saw tagsets that did not match the design).

One very prominent example of such a conceptual question is the partitioning of the part-of-speech space. Being an apparatus with 2300 years of history and tradition, parts of speech are defined based on a mixture of morphological, syntactic and semantic criteria. Moreover, their traditional grammatical definition, which is sometimes reflected in the tagsets, differs across languages. Thus what is called a determiner in one corpus may be tagged a pronoun in another, elsewhere even pronominal adverbs may be pronouns, yet elsewhere all pronouns and determiners may be dissolved in the classes of nouns and adjectives. It is technically not so difficult to add such features to Interset that all possible partitions of this space can be captured. However, we also want to make the Interset representation understandable by humans. The user should see both that a word has a pronominal function (e.g. demonstrative or interrogative) and that it behaves syntactically as a noun or an adjective. And it should be apparent even if it is not explicitly said in the source tagset but it can be deduced from it.

As for the normalization of the syntactic structures, we will have to first find a mechanism of identifying each type of structure in each treebank. We will use a combination of 1. parent-child links; 2. dependency labels (syntactic tags); 3. morphological tags. Sometimes we may even need to resort to the word forms. For instance, the Hindi treebank does not distinguish coordinating and subordinating conjunctions (neither morphological nor syntactic tags differ). However, conjunctions are a closed class and can easily be enumerated. If we see the very frequent word form कि *ki* ("that"), we know that we are not dealing with a coordinate structure.

We plan on using the Treex framework (Popel and Žabokrtský 2010) to implement all tree manipulations needed during the research.

The structural transformations that we will design should in general be reversible. Thus we will focus on clearly identifiable structures that are captured consistently across the original treebank (minus insignificant number of possible annotation errors). Examples include coordination, prepositional phrases, subordinate clauses or punctuation. We will be conservative not to damage language-specific phenomena where the original reasons for choosing a particular annotation style cannot be unambiguously identified. This holds both for the original normalization and for the later transformations that we will apply to assess learnability by parsers. We will also check the amount of non-projectively attached nodes in the original and in the normalized tree. Non-projective constructions are difficult for parsers and there should not be any reason to increase the number of non-projectivities during normalization.

Dependency labels will have to be changed whenever we reshape a structure in the dependency tree. Besides that, we will use one common set of syntactic tags. We will see whether it is possible (and to what extent) to apply the idea of Interset, without loss of information, also to syntactic tags. The skepticism is based on the observation that syntactic tagsets are *very* different across the treebanks, ranging from simple statements such as "this is a noun phrase modifying something" over standard *subject* and *object* relations to deep-level functions of Pāṇinian grammar in Indian languages.

We intend to use the annotation style of the Prague Dependency Treebank as the starting point for our normalized style. It will be gradually modified to accommodate constructions that are not present in PDT. The PDT style has already been applied to 10 languages from 3 families. At present we are not aware of any features in significantly different languages (such as Chinese) for which the PDT style could not be adapted or extended. Identification of such features (if they exist) will be an important outcome of our research.

Obviously the first step in adding a new treebank will be acquiring its documentation if possible. If it is not packed directly with the data, there might be descriptions online. Sometimes the documentation is not written in English. The proposer of this project speaks several languages and is able to decipher documentation in most Slavic, Germanic and Romance languages. If the documentation is incomplete or we cannot understand it, we will try to reach the creators for consultation. We will also automatically collect example occurrences of all part-of-speech, morphological and syntactic tags and study them in order to better comprehend their usage. To interpret the examples, we can use literature about the languages, on-line dictionaries and translation tools if necessary.

We are not convinced that it will be useful, with respect to the motivation of the work, to also normalize tokenization. Nevertheless we leave the possibility open and we will look at it during the project.

The experiments aimed at suitability of particular annotation styles for parsing will be conducted as follows. For a phenomenon (structure) that has K possible annotation styles we will define $2K$ transformation algorithms that will convert the structure from the normalized annotation style to the i-th style and back. A transformation algorithm will identify the nodes participating in the structure, reorganize and relabel them under the new style. Note that it will be much easier to have the normalized style as input, instead of the various original styles, because we will know exactly what to expect. Round-trip transformation to another style and back should be almost 100% lossless. We will train several different parsers on the transformed training data, apply the parsers to test data, transform the output back to the normalized style and evaluate parsing accuracy against the normalized gold standard annotation. We will compute the learnability scores for each transformation as defined by Schwartz et al. (2012). If an annotation style achieves the highest accuracy with most of the parsers, we can assert that the style is more suitable for parsing. Experimenting with different parsers (that are based on different machine learning approaches) will ensure that we will not bias the assertion towards one particular approach.

We also plan to follow up the previous work of (Zeman and Resnik 2008; McDonald et al. 2011; Täckström 2013) on cross-language parser adaptation. The basic idea is to use normalized morphological tags as features on which a parser is trained, then apply the parser to a new language. Such a technique can be useful to create a parser for a language for which no treebank exists, provided there is a treebank for a related language. We will naturally use the languages that do have a treebank so that we can evaluate the results on existing test data.

We want to make available to the research community as much of the resulting data as possible. About one third of the treebanks have licenses allowing us to freely modify and redistribute them. Most of the remaining corpora are freely available for research but the users have to obtain them directly from the original source. In these cases we can provide "patches" that can be combined with the original corpus. We will provide only our normalized annotation, without actually copying the underlying copyrighted text. In addition, all software that we will have to create in order to conduct the experiments will be freely available including source code, which should make any follow-up research easier. We also plan on releasing models for statistical taggers and parsers trained on the normalized corpora, so that researchers can machine-analyze previously unseen text.

Time Schedule

We will start with redesign of Interset, focusing on the issues mentioned above. We estimate the time needed to designing the new partitioning of the feature space, solving related conceptual questions and testing on a few very different tagsets, to 1–2 person-months (PM). Complete (re)writing and testing the conversion procedures between one tagset (language) and Interset could be achieved in 0.25 PM on average. The hardest and most time-consuming step will always be studying the data and documentation, collecting examples and understanding the original annotation guidelines. Since this has already been done for about a half of the treebanks, we decrease the total estimate from $30 \times 0.25 = 7.5$ PM to about 5 PM.

Syntactic normalization can be divided to an easy and a difficult part. Conceptually easy is for instance the inner structure of prepositional phrases, subordinating conjunctions, attachment of some punctuation (paired quotation marks and parentheses, sentence-terminating punctuation). We have discussed these phenomena previously during our initial work towards HamleDT, and we believe that 0.125 PM per language will be enough to make sure that all the languages have these structures normalized correctly.

The difficult part includes coordinate structures and various verbal groups (compound verb forms, participles + auxiliaries, modals + infinitives etc.) Both involve potentially many tree nodes and both require extensive investigation of the existing annotation styles and careful design of transformation algorithms. We estimate 1–2 PM for the general design of the algorithms. Once these are ready, processing of one language may again take about 0.125 PM. (Occasional feedback to the overall design may be needed; it is included in the time estimated for the design.)

Most of the work will be done by two researchers. We will allocate 40–50 % of their working capacity to the project (i.e. 1 PM equals to 2–2.5 absolute months of one researcher's work; see also the Justification of Personal Costs). One researcher will be prevailingly responsible for investigating the original annotation styles and normalization (research line 1, RL1), while the other will design the transformations between normalized and other styles, train taggers and parsers, run and evaluate parsing experiments (RL2). We will not first normalize all the languages, then run the parsing experiments. Instead, these two working packages will run concurrently for the most part of the project. There are two reasons for this decision: 1. By doing the two subtasks concurrently we hope to earlier discover potential issues resulting from their interaction. 2. Having one person responsible for normalization of all the languages will help maintain the same standard for all the languages.

During the first four months of RL1, Interset will be redesigned and rewritten. Then we will gradually write the new Interset conversion procedures for all the languages. We will start with the languages where the initial exploration has already been done so that first datasets are available quickly for RL2. Together with Interset, we will also prepare syntactic normalization of the "easy" phenomena and of coordinate structures (which are not "easy" but we cannot work with the rest of the tree without normalizing coordination first). By the end of year 1 we should have about two thirds of the languages ready.

RL2 in the first half of year 1 will involve obtaining and training taggers for all the languages (relatively independent of Interset and needed later to be able to process unseen text), preparing the Treex infrastructure

for transformations, conceptual design especially for coordinate structures. As the first newly normalized corpora will become available, transformations of coordinate structures will be thoroughly tested, especially with respect to the lossless roundtrip conversion. (One could also think of experimenting with the previous prototype of HamleDT before the new normalization is available. Unfortunately our previous experience shows that it is difficult to draw any conclusions before all remaining flaws in normalization are corrected.) In the second half of year 1, RL2 will add transformations of the "easy" syntactic phenomena mentioned above.

We reserve one month of each year for work on journal articles and conference papers, preparing conference presentations and attending the conferences (this is not a person-month; our estimate is that every team member will on average spend one month on this sort of work, according to their working capacity: a person dedicating 50 % of their working capacity to the project will thus spend half a month on this). We further reserve one month of each year for the legal holiday (25 working days each employee).

In the year 2, RL1 will first finish the Interset normalization + "easy" and coordination normalization for the remaining languages. Then the "difficult" phenomena (esp. verb groups) will be investigated, their normalization designed and immediately tested on one third of the languages. In the middle of the year, this normalization will be gradually applied to the remaining languages. If everything goes smoothly, there should be one to two months reserve at the end of the year. We will start preparing the book at this time so that it can be sent to the reviewers in the second half of year 3.

RL2 in year 2 will focus on experiments with cross-language parser adaptation. At the time of designing normalization for the "difficult" phenomena, RL1 and 2 will interact to get pilot transformations and feedback. However, the full-scale learnability experiments with these structures will be done in year 3. RL2 will also explore possibilities of unifying tokenization approaches in all the treebanks, and evaluate the merit.

In the year 3, the main responsibility of the researcher in RL1 will be preparing the book, making good use of the experiences from the normalization work. Naturally there will be input from the other members of the team, too. Besides that, RL2 will repeat previous parsing experiments with the final data ("difficult" stuff included). The monograph will be sent to reviewers at the end of summer. We (RL1 and 2 combined) plan on using the remaining time after that point to add a few extra languages, possibly also adding other treebanks of existing languages. A non-trivial time will also be needed to prepare the normalized annotations in a form that could be distributed to other research teams.

The Institute and the Team Members

The Institute of Formal and Applied Linguistics (ÚFAL) has long-standing experience with natural language text processing and linguistic data (corpora) preparation. In this field it belongs to the national elite and is accepted also worldwide. It was founded in the beginning of the 1990s as a follow-up of several decades of the so-called Prague Linguistic School, of pedagogical work in formal linguistics at the Charles University (Univerzita Karlova) in the second half of the 20th century. Among the most notable results of the institute is the Prague Dependency Treebank, a set of linguistically annotated data of the Czech language, unique both by its extent and its depth of processing. Several other large-scale treebanks have followed, for example the Prague Czech-English Dependency Treebank (PCEDT), Prague Arabic Dependency Treebank (PADT) or Prague Dependency Treebank of Spoken Language (PDTSL).

The members of the institute have worked on many other natural language processing projects, ranging from morphological analysis and tagging over syntactic parsing, valency, topic-focus articulation, and text generation to search in large text collections and machine translation. The institute collaborates with many top-class research institutions abroad, e.g. the Center for Language and Speech Processing at JHU Baltimore (MALACH, machine translation, summer workshops etc.), University of Maryland Institute for Advanced Computer Studies (MALACH, information retrieval, machine translation), Dublin City University (machine translation) or IRCS and LDC at the University of Pennsylvania, Philadelphia (corpora, grammar, machine translation, Arabic treebank). As a member of LDC and ELDA, the institute has access to a large scale of linguistic data. In particular, the institute already has got research usage rights to all the treebanks that we plan on investigating.

The institute is also well prepared to processing of huge quantities of data (such as repeated training of *N* parsers on 30 treebanks times *K* annotation styles). ÚFAL's server facilities consist of a data cluster and a computational cluster with the overall computing capacity of 600 processor cores, 4TB of memory and 40TB

of disk storage. The cluster is interconnected by a 10Gbps backbone. It supports the creation of dynamic Hadoop clusters to optimize the parallelization for processed tasks.

Team Members

RNDr. **Daniel Zeman**, Ph.D. has been working many years on syntactic parsing of Czech sentences using statistical methods. In 2005 he defended his dissertation ("Parsing with a Statistical Dependency Model") (Zeman 2004). He took part in several international projects, in 1998 the summer workshop at the Johns Hopkins University, Baltimore, about the adaptation of various methods to parsing of English and Czech; in 1999 University of Pennsylvania, Philadelphia, the topic was designing a method of automatic extraction of subcategorization frames from a corpus (Sarkar and Zeman 2000); Czech was the model language but the publication got broad international response from teams working on Greek, Chinese and other languages. In 2006 he was awarded a one-year scholarship of the J.W. Fulbright Commission at the University of Maryland, College Park. He worked on cross-language parser adaptation, and started collaboration with the team of Dr. Resnik; Interset was a by-product of this project. Recently, Dr. Zeman was the principal investigator of another GAČR project called CZECHMATE ("Czech in the Machine Translation Era"). He also works as a reviewer for main international conferences and journals and teaches two NLP courses at the Charles University and the Czech Technical University.

Mgr. **Martin Popel** is a PhD candidate and research assistant at ÚFAL who graduated in 2009 with the thesis "Ways to Improve the Quality of English-Czech Machine Translation". His main research interests are dependency-based machine translation, machine learning and parsing. Since 2008, he works on the TectoMT deep-syntactic machine translation system that is achieving competitive results with state-of-the-art systems in the yearly WMT evaluation campaigns. Since 2011 he is the main developer of the Treex NLP software framework. He has been working on three EU-funded projects, EuroMatrixPlus, Khresmoi and QTLeap. He teaches a course on modern methods in computational linguistics at the Charles University.

Doc. Ing. **Zdeněk Žabokrtský**, Ph.D. works in the field of natural language processing since 2000. He defended his PhD thesis (Valency Lexicon of Czech Verbs) in 2005, and in 2011 he became an associate professor in mathematical linguistics. He supervises several PhD students and teaches two courses on exploiting language data and two courses on machine learning. His main research areas are syntax-based machine translation, building language data resources, dependency syntax, coreference resolution, and applications of machine learning in natural language processing. He could contribute to the proposed project especially with his expertise in dependency syntax and multilingual data processing, which he has gathered in a number of projects in the past. He has participated in the development of the Prague Dependency Treebank and the Prague Czech-English Dependency Treebank, and contributed to development of treebanks for some other languages (e.g. Slovene and Tamil). He studied transformations between different syntactic formalisms, and implemented dependency parsers for several languages (e.g. Romanian, Polish, Arabic, German) before dependency treebanks were available for them. He is one of the main authors of the Czech-English parallel treebank CzEng, which is now intensively used for machine translation experiments. He has supervised his student's work on building the W2C corpus, which is a unique collection of web-based corpora for about 100 languages.

Dr. Zeman will be responsible for RL1 (see Time Schedule), Mgr. Popel for RL2. Doc. Žabokrtský will participate in the strategical discussions and design work; during the whole duration of project he will be providing input through consultations and he will take part in preparing the publications. We plan to allocate 10% of his working capacity to the project.

We strongly believe that our previous experience warrants appropriate qualification needed for this kind of research. Dr. Zeman is the author of the original Interset, he has worked on natural language processing of numerous languages (parsing of dozens of languages from Czech through Arabic to Hindi; morphemic segmentation of Turkish and Finnish; machine translation between multiple European and Indian languages etc.) Doc. Žabokrtský and Mgr. Popel are leading developers of the Treex framework that we intend to use to implement the transformations. All three team members participated in formulating the first prototype of HamleDT.

Expected International Collaboration

We have collaborated with treebank designers abroad in the past. Doc. Žabokrtský recently provided technical and mentoring assistance to the team of the (Danish) Copenhagen Dependency Treebank; we have been exchanging knowledge with the linguistic department at IIIT Hyderabad (treebanks of Hindi, Urdu, Bengali and Telugu); similarly, we have been in touch with prof. Nivre (Uppsala University) and dr. McDonald (Google), world-renowned experts on dependency parsing and treebanks. We will reuse these contacts (and possibly add new ones) to get feedback and extend our know-how.

References

Bengoetxea K. and Gojenola K. (2009) Exploring Treebank Transformations in Dependency Parsing. In *Proceedings of RANLP 2009*, Borovec, Bulgaria.

Bosco C., Montemagni S., Mazzei A., Lombardo V., Lenci A., Lesmo L., Attardi G., Simi M., Lavelli A., Hall J., Nilsson J. and Nivre J. (2010) Comparing the Influence of Different Treebank Annotations on Dependency Parsing. In *Proceedings of LREC 2010*, Valletta, Malta.

EAGLES (1996) *Recommendations for the Morphosyntactic Annotation of Corpora*. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html

Intercorp. *Český národní korpus – InterCorp.* Ústav Českého národního korpusu FF UK, Praha. Cit. 11.4.2013, http://www.korpus.cz/

Johansson R. and Nugues P. (2007) Extended Constituent-to-Dependency Conversion for English. In *Proceedings of NODALIDA 2007*. Tartu, Estonia.

McDonald R., Petrov S. and Hall K. (2011) Multi-source Transfer of Delexicalized Dependency Parsers. In *Proceedings of EMNLP 2011*, Edinburgh, Scotland.

McDonald R. et al. (2013) Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL 2013*, Sofia, Bulgaria.

Nilsson J., Nivre J. and Hall J. (2006) Graph Transformations in Data-Driven Dependency Parsing. In *Proceedings of ACL-COLING*, Sydney, Austraila.

Popel M. and Žabokrtský Z. (2010) TectoMT: Modular NLP Framework. In *Proceedings of IceTAL*. Reykjavík, Iceland.

Sarkar A., Zeman D. (2000) Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of COLING 2000*, Saarbrücken, Germany.

Schwartz R., Abend O. and Rappoport A. (2012) Learnability-Based Syntactic Annotation Design. In *Proceedings of COLING 2012*, Mumbaī, India.

Sedláček R. and Smrž P. (2001) A New Czech Morphological Analyser ajka. In *Proceedings of TSD 2001*. Berlin / Heidelberg, Germany: Springer.

Täckström O. (2013) *Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision.* PhD thesis, Uppsala universitet, Uppsala, Sweden.

TEI Consortium, ed. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/

Tsarfaty R., Nivre J. and Andersson E. (2011) Evaluating Dependency Parsing: Robust and Heuristics-Free Cross-Annotation Evaluation. In *Proceedings of EMNLP 2011*, Edinburgh, Scotland.

Windhouver M.A. (2012) Towards Standardized Descriptions of Linguistic Features: ISOcat and Procedures for Using Common Data Categories. At the KONVENS 2012 workshop *Standards for Language Resources*. Wien, Austria.

Zeman D. (2004) *Parsing with a Statistical Dependency Model.* PhD thesis, Univerzita Karlova v Praze.

Zeman D. (2008) Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Zeman D., Mareček D., Popel M., Ramasamy L., Štěpánek J., Žabokrtský Z. and Hajič J. (2012) HamleDT: To Parse or Not to Parse? In *Proceedings of LREC 2012*, İstanbul, Turkey.

Zeman D. and Resnik P. (2008) Cross-Language Parser Adaptation between Related Languages. In *Proc. of IJCNLP 2008 workshop on NLP for Less Privileged Languages*. Hyderabad, India.