The Current Status of the Prague Dependency Treebank

Eva Hajičová[12], Jan Hajič[12], Barbora Hladká[1], Martin Holub[1],
Petr Pajas[1], Veronika Řezníčková[12], and Petr Sgall[1]

[1] Center for Computational Linguistics, Charles University (MFF),
Malostranské nám. 25, CZ-11800 Prague, Czech Republic,
{hajicova,hajic,hladka,holub,pajas,vr,sgall}@ckl.mff.cuni.cz,
http://ckl.mff.cuni.cz

[2] Institute of Formal and Applied Linguistics, Charles University (MFF),
Malostranské nám. 25, CZ-11800 Prague, Czech Republic,
{hajicova,hajic,vr}@ufal.mff.cuni.cz,
http://ufal.mff.cuni.cz

Abstract. The Prague Dependency Treebank (PDT) project is conceived of as a many-layered scenario, both from the point of view of the stratal annotation scheme, from the division-of-labor point of view, and with regard to the level of detail captured at the highest, tectogrammatical layer. The following aspects of the present status of the PDT are discussed in detail: the now-available PDT version 1.0, annotated manually at the morphemic and analytic layers, including the recent experience with post-annotation checking; the ongoing effort of tectogrammatical layer annotation, with a specific attention to the so-called model collection; and to two different areas of exploitation of the PDT, for linguistic research purposes and for information retrieval application purposes.

1 Introduction

It is our conviction that an existence of a large corpus together with a rich annotation scheme applied to it offers a quite new level of possible topics for investigation, using the annotated data themselves or data gained by automatic tagging procedures developed on their basis.

Therefore, in the build-up of the Prague Dependency Treebank (PDT), for which we use a subcollection of texts from the Czech National Corpus ([4]) we have tried to develop a scenario that would be as multi-aspectual as possible. This is reflected first of all in the overall annotation scheme conceived of as a three-layer scenario comprising tags from the morphemic, analytic and underlying-syntactic layer (for a description of the annotation scheme of PDT, see e.g. [5], [8], [6], [10], [11], and the two manuals for annotation published as Technical Reports ([7], [12]).

The annotation on the highest, underlying-syntactic layer, the result of which are the so-called tectogrammatical tree structures (TGTS) is based on the original theoretical framework of Functional Generative Description as proposed by

Petr Sgall in the late sixties and developed since then by the members of his research team ([19]). It goes without saying that such a rich description of the morphemic and syntactic properties of sentences (including some basic coreferential relations) cannot be achieved without a thorough and detailed inspection of the language corpus itself (before we can attempt to work on an automatic tagging procedures, be it stochastic or based on hand-written rules). Therefore, we have annotated each and every sentence in the PDT manually, with the help of several productivity software tools developed during the project (which, however, did include some automatic preprocessing modules).

In the present contribution we would like to discuss in some detail the present status of the development of the PDT and the experience we have hitherto gained. In Sect. 2, we describe the just finished annotation of the morphemic and analytic layers of PDT (which we call PDT version 1.0), discussing in some length our experience with the post-annotation checking; our first experience with the annotation on the tectogrammatical layer is summarized in Sect. 3. Sect. 4 then presents some examples of possibilities the tagged corpus offers for researchers in both linguistics and natural language processing applications (specifically, in information retrieval). We conclude with some observations concerning the outlook of the PDT.

2 The PDT, Version 1.0

2.1 PDT 1.0 Overview

Since the process of manual annotation of tens of thousands of sentences is a lengthy one, and since we want to make our results available to the community promptly, we have decided not to wait until all three layers are annotated, but to release the annotation on the first two layers (morphemic and analytic) right after it is finished. We call the result the PDT, version 1.0 ([16])[1]. The annotation consists of a unique (*lemma*, *tag*) pair at the morphemic layer and a unique (*head pointer*, *analytical function*) pair at the analytic layer assigned to each token (word form, number, or punctuation occurrence) in the corpus.

The PDT 1.0's data layout facilitates experiments based on various methods (primarily statistical, but not only those) and especially allows for their fair comparison. The data on each layer is thus already divided into a training set and two sets of evaluation data. The morphological tagging as part of the analytic layer annotation has been provided by two different statistical taggers ([8], [9]).

About 1.8 million tokens have been annotated on the morphological layer, and 1.5 million tokens (almost 100,000 sentences) on the analytic layer. The data itself is marked using a common SGML DTD (`csts.dtd`), and a different format is used for some viewing and editing by legacy tools provided for PDT users.

The organization of such an annotation effort is not an easy task. A total of 32 people contributed to this project to this date, with as many as 20 working simultaneously at a peak time.

[1] The "version 0.5" (ca. 1/4 of the data annotated on the two layers) has been available since 1998 on our website and has attracted 90 researchers from 17 countries.

2.2 Post-Annotation Checking

In a manually annotated corpus, the single most important issue is *consistency*. It is thus natural to devote a large proportion of time (*resources*) to corpus checking, in addition to the annotation proper.

There was no annotation manual on the morphemic layer (the categories in the tagset used correspond directly to what every high school graduate knows about Czech morphology), but the data was double-tagged as usual. While nine annotators have been involved in the first pass of the annotation, only two annotators have participated in the checking step. The discrepancies coming from the two annotated versions of the same file were checked and disambiguated by only one annotator. After that, each file has been checked against the automatic morphological analyser (AMA), which produces a set of (*lemma*, *tag*) pairs for each word form. If the manually assigned pair is not found in this set, we have (a relatively easily solvable) consistency problem: either the AMA is wrong or incomplete, or the manual annotation has to be corrected by manual inspection. The annotation problems found here are, however, not just plain annotation errors; misspellings in the original text and formatting errors (such as wrongly split or joined word forms) are discovered and corrected, keeping the original input (properly marked) for further exploitation, such as spelling error analysis. The AMA check had to be done several times, since the dictionary of the AMA has changed several times during the time frame of the project.

The situation with the analytic layer was a bit different. The group of annotators[2] (being all linguists by education) has been writing a common set of Guidelines ([7]) during the course of annotation (most of it at the beginning of the project, of course), solving new problems on-the-go. Needless to say that the solutions of some of the problems affected the already-annotated part of the corpus, leading to re-annotation of certain phenomena in it. However careful the annotators have been when doing so, inconsistencies could not be avoided. Moreover, our limited resources forced us to annotate every text by one annotator only, and we have used some automatic processing during the annotation process as well: the analytical functions have been preassigned by a small set of hand-crafted rules, which operated on the manually created sentence structure, and during the later stages of the annotation process, we have also preassigned the sentence structure using a statistical parser ([3]) trained on data annotated so far. In both cases, the annotators have been instructed to correct both the structure and the analytical functions to conform to the Guidelines. Once the annotation of all the data at the analytic layer had been finished, we have applied a list of 51 consistency rules ("tests" regarding the linguistic content), created by inspection of the data and by formal specification of known problems. The tests were intended to help us locate the most evident mistakes that the annotators, authors or programs could have made during the process of annotation. Some of those test (and corrections) could be done almost fully automatically, but some of them had to be carried out manually (after automatic preparation and flagging of questionable spots). Several additional manual (even though sometimes

[2] A different group than that for the morphemic layer.

only partial) passes through the data were thus required. Additional tests have been designed and carried out to discover technical problems, such as missing or incorrect markup etc.

Since the PDT 1.0 contains two layers (morphemic and analytic), we could take advantage of the relations between the two layers for checking purposes as well, improving consistency *across* the layers at the same time. The sentence context of the analytic layer allowed us to discover otherwise hidden annotation errors on the morphemic layer and vice versa[3]. The following is a partial list of the cross-layer checks made:

1. The verb complements at the analytic layer (namely the nodes annotated by the analytical function `Obj` (object), `Sb` (subject) and `Pred` (predicate)) were tested against their morphological tags.
2. Prepositional phrases (PPs) headed by certain prepositions listed in the Guidelines have been checked for the analytical function against those lists. The case(s) of nouns, pronouns, numerals, and adjectives inside a PP have been checked against the possible "valency" of its head preposition.
3. Agreement in case, gender and number between a predicate and its subject (`Sb`), as well as between and attribute (`Atr`) and its head was checked.

The *resources* needed for the annotation on the morphemic and analytic layers can be roughly estimated as follows (in percent of the total manpower):

1. the "raw" morphemic layer annotation: 25%
2. the "raw" analytic layer annotation: 15%
3. post-annotation checking (both layers), related software development: 20%
4. data processing (layer merging) and associated manual corrections: 5%
5. documentation (incl. analytic-layer Guidelines): 5%
6. annotation software tool development: 25%
7. supervision and administration (both layers): 5%

The total manpower does *not* include the development of the morphological analyzer and the morphological dictionary of Czech, the Czech taggers, the Czech version of the analytic (syntactic) parser, nor the effort needed for the initial collection and basic markup of the texts used.

3 Towards PDT 2.0: the Tectogrammatical Layer

The tectogrammatical annotation ([10], [11], [12]) of the PDT is carried out on two sub-levels, resulting in (i) a 'large' collection which captures the underlying syntactic structure of the sentence (in the shape of dependency tree structures distinguishing about 40 types of syntactic relations called functors) and the basic features of the topic-focus articulation (TFA) in terms of three values of

[3] Of course, manual correction had to be done after the suspicious annotations have been flagged by the checking software.

a TFA attribute and of the underlying order of sister nodes of each elementary subtree of the dependency tree, and (ii) a 'model' collection with more subtle distinctions of valency relations (in terms of a subcategorization of the valency slots by means of the so-called syntactic grammatemes primarily capturing the meanings of prepositions) and with values that indicate the basic coreference links between nodes (within the same sentence but also across sentences).

In the present contribution we illustrate the complexity of the task of the annotation of the 'model' collection on some of the issues concerning the restoration of nodes for semantically obligatory complementations (valency slots) of verbs and of postverbal nouns and adjectives and those concerning coreferential relations of the restored nodes to their antecedents.

At the present stage, the annotators have restored semantically obligatory complementations as dependents of the given verbs, postverbal nouns or adjectives according to the following instructions:

(i) restore a node with the lemma Gen in case a General Participant is concerned: e.g. in the TGTS for the sentence Náš chlapec už čte. [Our boy already reads.] a node is restored depending on the verb číst [read] with the lexical label Gen and the functor PAT; the attribute Coref for coreference is left untouched;

(ii) restore a node with the lemma Cor in case of grammatical coreference (i.e. with verbs of control, with relative pronouns and the possessive pronoun svůj): e.g. in the TGTS for the sentence Podnik hodlá zvýšit výrobu.[The company intends to increase the production] a node is restored depending on the verb zvýšit [increase] with the lexical label Cor and the functor ACT; the attribute of coreference gets the relevant values;

(iii) restore a node with the pronominal lemma on in case of textual coreference (i.e. the deletion of the respective node in the surface shape of the sentence is conditioned by the preceding context rather than by some grammatically determined conditions): e.g. in the sequence of sentences Potkal jsi Jirku? Potkal. [Have you met Jirka? (I-)Met.] two nodes are restored in the TGTS of the second sentence depending on the verb potkat [meet], one with the pronominal lemma já and the functor ACT and one with the pronominal lemma on and the functor PAT; with the latter node the attribute of Coref is filled in by the lemma Jirka.

Our experience with the first samples of sentences tagged for the 'model' collection has shown that for the restoration of obligatory participants with verbs and postverbal nouns and adjectives in cases of textual coreference a new lemma Unsp(ecified) has to be introduced in order to capture situations when the restored node refers to the 'contents' of the preceding text rather than to some particular element; the information on the antecedent is vague. On the other hand, the restored lemma differs from the lemma Gen introduced for General Participants because, in principle, with Gen no antecedent is present . The attribute of Coref with the restored node has the value NA (=non-applicable). We believe that this solution offers a possibility of further linguistic inquiries into the issues of coreferential relations because it leaves a trace specifying the problematic cases.

Let us illustrate the above points on some examples from the PDT. We add literal English translation for each sentence.

(1) Prudký růst maďarského průmyslu.

(2) Maďarská průmyslová výroba se v loňském roce zvýšila o devět procent v porovnání s rokem 1993.

(3) Ve stavebnictví byl zaznamenán dokonce dvacetiprocentní přírůstek.

(4) Vyplývá to z údajů, které v pátek zveřejnil centrální statistický úřad.

(5) Spotřební ceny stouply v meziročním srovnání o 18,8 procenta, zatímco v roce 1993 dosáhla míra inflace 22,5 procenta.

(1') A rapid increase of Hungarian industry.

(2') Hungarian industrial production increased in the last year by nine percent in comparison with the year 1993.

(3') In building industry (there) was recorded even a twenty percent increase.

(4') (it) follows from the data which on Friday (were) published (by) the National Census Bureau.

(5') The prices went up in the yearly comparison by 18.8 percent, while in the year 1993 the rate of inflation reached 22.5 percent.

In sentences (1) through (5) there occur several instances of verbs and postverbal nouns the complementations of which have to be restored. In (3) and (5) the nodes for a General Actor (Gen.ACT) have to be restored as dependents on the verb zaznamenat [record] and on the noun srovnání [comparison], respectively. The lexical label Gen indicates that no antecedent with these nodes exists, because the Actors can be paraphrased as 'those who recorded' and 'those who compared', respectively. However, the second obligatory complementation, namely the Patient (PAT) in the frame of the postverbal noun srovnání [comparison] can be uniquely determined: it is clear from the context that a comparison of prices is concerned. The restored node for the Patient gets the pronominal lemma on and the attribute Coref gets the value cena [price]. A similar character has the restored node for the Actor dependent on the postverbal noun přírůstek [increase] in (3). It is clear from the preceding sentence (2) that the Hungarian industrial production increased, i.e. that an increase of production is concerned. Therefore the restored node for Actor depending on přírůstek [increase] gets the lemma on and the attribute Coref gets the value výroba [production]. On the other hand, in the TGTS of (4) a node for the Patient depending on the verb zveřejnit [publish] should be restored (the valency frame of this verb can be paraphrased by 'someone.ACT publishes something.EFF about something.PAT'), but its antecedent is rather vague: we can guess from the context that the statistical institute will publish the data on the problems of the increase of the Hungarian industry, but the increase may concern the Hungarian industrial production or the building industry - the concrete reference is not clear. Thus the restored node gets the lemma Unsp and the attribute Coref gets the value NA.

Let us add two more examples illustrating the restoration of nodes with the lemma Unsp:

(6) Poslanecká sněmovna schválila novelu zákona o mimosoudních rehabilitacích.

(7) Novela, jíž nakonec český parlament posvětil, má však tolik zádrhelů, že k jásotu není sebemenší důvod.

(8) Za jednoznačné pozitivum lze považovat snad jen fakt, že zákon vůbec prošel.

(9) Jeho schválení předcházela úporná jednání uvnitř koalice.

(6') The House of Commons approved the novel of the law on out-of-court rehabilitations.

(7') However, the novel, which in the end the Czech Parliament has sanctioned, has so many trouble spots that for jubilation (there) is not the slightest reason.

(8') As clear positive can be considered perhaps only the fact that the law has been approved.

(9') Its approval (was) preceded (by) tough negotiations within he coalition.

In the surface shape of (9) there is no overt Actor with the postverbal noun schválení [approval] ; the restored Actor will get the lemma Unsp because it is not clear from the context whether the restored node refers back to the House of Commons, or the Parliament. As for the postverbal noun jednání [negotiations], all the restored dependents will get the lemma Unsp: though it is possible to guess that some coalition party (ACT) negotiated with some other coalition party (ADDR) about the problems (PAT) of the novel of the law, no coalition parties have been mentioned in the previous context and the reference again is only vague. The attribute of Coref with all the three restored nodes will get the value NA.

(10) Páteční rozhodnutí sněmovny , že ... "předjímá svým způsobem další vývoj diskusí o církevních restitucích".

(11) Zpravodaji LN to včera řekl místopředseda parlamentu Jan Kasal (KDU - ČSL).

(12) Spolu s precedentním pátečním zásahem do obecního majetku tak podle něho mizí velká část dosavadních překážek.

(10') Friday decision of the House of Commons that ... "(it) anticipates in a way the further development of the discussions on the church restitutions."

(11') The correspondent of LN (was) told this (by) the vice-chairman of the Parliament Jan Kasal (KDU-CSL).

(12') Together with the precedent Friday intervention into the communal property thus according to him disappears a great part of the hitherto obstacles.

In (10) an Actor is restored depending on the noun diskuse [discussion], with the lemma Unsp (someone.ACT discusses with somebody.ADDR about something.PAT); it is probable from the context that the discussion will be carried out by the political parties in the Parliament, but again this cannot be determined univocally. In (12) an Actor should be restored under the postverbal noun zásah [intervention], with Unsp as its lexical label because it is not clear whether the intervention was made by the House of Commons, or whether the speaker has somebody else in mind.

The decision on the boundary lines among the different types of deletions linked to the choice among the “lemmas” of the restored nodes is not be an easy task. Our strategy again is to mark the difficult cases in a way that allows for their relatively simple identification and thus to prepare resources for further linguistic research.

4 Exploitation of the PDT

4.1 Linguistic Research

From the very beginning, the annotation of the PDT has been guided by the effort not to loose any important piece of information encoded in the text itself, but at the same time not to overload the annotation scheme and thus not to prevent the annotators to present some reliable and more or less uniform results. It is then no wonder that the most obvious exploitation of the PDT is for linguistic research as such, which, in its turn, offers most important material for the improvement, precisation and clarification of the annotation instructions, and, in the long run, for possible modifications of the scheme itself. As an example, let us mention the research on the so-called PP-attachment (i.e. the ambiguities resulting from the possibility to attach a prepositional group to more than one of the preceding elements) carried out by Straňáková-Lopatková ([21]) . The author formulated an algorithm for the solution of these ambiguities based on an original formal framework of deletion automata ([17]) and she used the PDT both as an empirical basis for her study and as a testing bed.

4.2 Annotation Automation

Another crucial point of an annotation scenario is the division of labor between automatic and manual procedures; one would prefer as much automation as possible while not compromising the precision of the human annotation. Such considerations have led to a development of an experimental system for automatic functor (pre-)assignment which is based on a machine-learning approach ([22]). Possibilities are examined how to resolve ambiguities of functor assignment on the basis of the meanings of prepositions and their combinability with nouns of different positions in the EuroWordNet ontology. We are also building a valency dictionary containing information relevant on all layers, i.e. functors (for TGTS) and morphosyntactic information, on the basis of valency frames from various sources, primarily from the material already contained in the PDT.

4.3 Applications: Information Retrieval

Searching for information in huge amounts of full texts is still mostly word-based. This technique is often ineffective because the matching process based on word forms does not respect natural language. To avoid this problem one can try and match concepts expressed by word forms rather than words themselves.

When we want to work with concepts instead of words, the information retrieval system should be able to compare the concepts in order to determine (or at least to estimate) their semantic similarity or differences.

5 Conclusion

We have tried to demonstrate on some selected issues how complex the task of syntactic annotation of a corpus is and what solutions have been chosen to make it usable. The future efforts will be concentrated on four domains: (i) to reach a solid volume of annotated data on the third layer, (ii) to extend the scenario to make it possible to design some kind of a formal semantic (logical) representation (the "fourth" layer), and eventually to annotate such layer of the PDT, (iii) to use the PDT in the domains of information retrieval, information extraction and/or computerized translation, for the purpose of which an extensive work with parallel (or at least comparable) corpora is a necessary precondition, and (iv) to prepare grounds for a similarly systematic compilation and annotation of a spoken language (speech) corpus. Ars longa, vita brevis.

6 Acknowledgments

This work has been supported by various grants and project. The main contributors are the Grant Agency of the Czech Republic, project 405/96/K214, and the Ministry of Education, project Center of Computational Linguistics (project LN00A063).

References

1. Böhmová A., Hajič J., Hajičová E., Hladká B.: The Prague Dependency Treebank: A Three-Level Scenario. In: Treebanks: Building and Using Syntactically Annotated Corpora, ed. Anne Abeille, Kluwer Academic Publishers, in press
2. Böhmová A., Panevová J., Sgall P.: Syntactic Tagging. In: Text, Speech and Dialogue. Ed. By V. Matoušek, P. Mautner, J. Ocelíková and P. Sojka. Berlin:Springer (1999) 34–38
3. Collins M., Hajič J., Brill E., Ramshaw L., Tillmann Ch.: A Statistical Parser of Czech. In Proceedings of 37th ACL'99 (1999) 22–25
4. Czech National Corpus on-line resources: `http://ucnk.ff.cuni.cz`
5. Hajič J.: Building a Syntactically Annotated Corpus In: Issues of Valency and Meaning, ed. by E. Hajičová. Prague: Charles University (1998) 106–132
6. Hajič J.: Disambiguation of Rich Inflection - Computational Morphology of Czech. Vol. I. Prague: Karolinum, Charles University Press (2001) 334pp.
7. Hajič J. et al.: A Manual for Analytic Layer Tagging of the Prague Dependency Treebank (1999). English translation. Technical Report, UFAL MFF UK. In prep.
8. Hajič J., Hladká B.: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings of Coling/ACL'98. Montréal, Canada (1998) 483–490

9. Hajič, J., Krbec, P., Květoň, P., Oliva, K., Petkevič, V.: Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In: Proceedings of the 39th ACL Meeting, Toulouse, France (2001). In print.
10. Hajičová E.: The Prague Dependency Treebank: From Analytic to Tectogrammatical Annotations. In: Text, Speech, Dialogue, ed. by P. Sojka, V. Matoušek and I. Kopeček, Brno: Masaryk University (1998) 45–50
11. Hajičová E.: The Prague Dependency Treebank: Crossing the Sentence Boundary. In: Text, Speech and Dialogue, ed. by V. Matoušek, P. Mautner, J. Ocelíková and P. Sojka, Berlin: Springer (1999) 20–27
12. Hajičová E., Panevová J., Sgall P.: A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. ÚFAL/CKL Technical Report TR-2000-09, Charles University, Czech Republic (2000)
13. Panevová J.: On Verbal Frames in Functional Generative Description. Prague Bulletin of Mathematical Lingusitics 22, 3-40, 23, (1974) 17–52
14. Panevová J.: Formy a funkce ve stavbě české věty (Forms and Functions in the Sentence Structure of Czech). Prague: Academia (1980)
15. Panevová J.: Ellipsis and Zero Elements in the Structure of the Sentence. In: Tipologie, grammatika, semantika. Sankt-Peterburg: Nauka (1998) 67–76
16. PDT on-line resources: `http://ufal.mff.cuni.cz/pdt`
17. Plátek, M.: Strict Monotonicity and Restarting Automata. PBML 72 (1999) 11–27
18. Řezníčková V.: PDT: Two Steps in Tectogrammatical Syntactic Annotation. (2001) Delivered at the SLE Annual Meeting, Leuven
19. Sgall P., Hajičová E., Panevová J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: Reidel (1986)
20. Skoumalová H., Straňáková M., Žabokrtský Z.: Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. This Volume.
21. Straňáková M.: Ambiguity of Prepositional Groups and the Possibility of Its Automatic Processing. PhD Thesis, Charles University, Prague (2001)
22. Žabokrtský Z.: Automatic Functor Assignment in the Prague Dependency Treebank. Master Thesis. Czech Technical University, Prague (2000)