Large Language Data and the Degrees of Automation

Alena Böhmová, Eva Hajičová

{bohmova,hajicova}@ufal.mff.cuni.cz
Center of Computational Linguistics, Faculty of Mathematics and Physics
Charles University Prague

1. Introduction.

Annotation of very large corpora is a complex and multifarious task, in the process of which the annotating teams are confronted with various aspects of linguistics and information science. In the present paper, we discuss the attempts to balance the degree of automatic processing, the accuracy of the annotation and the requirements laid by the amount of annotated data as realized in the course of the annotation of the Prague Dependency Treebank (PDT). We will focus on some points connected with the issues of the depth of annotation, of the robustness of the annotating scheme and the coverage.

2. The depth, robustness and coverage of annotation

The requirements of the depth and precision of annotations vary for different intended uses of the corpus. While the statistical language modelling methods are greedy for data and less demanding for accuracy, we also need to consider linguistic uses of the corpus - which contrive with much smaller amounts of data but depend on its 'total' accurracy. The multi-layered scenario of PDT takes this scale into account.

First, let us consider the scale of the depth of annotation. The PDT project is well documented from this point of view in publications of the research team (Böhmová, Hajič, Hajičová, Vidová-Hladká 2003 , Hajič, Hajičová, Holub, Řezníčková, Pajas, Sgall and Vidová-Hladká 2001, Hajičová 1999, Hajičová 2002), therefore we mention only the main aspects here. The annotation scheme has been elaborated based on the Praguian theory of Functional Generative Description (for a detailed description, see Sgall et al.1986), aiming towards the deep syntactic (tectogrammatical) level of description. Starting with morphemic analysis and morphemic tagging (disambiguation, see Hladká 2000, Hajič 2003), with about 1100 different tags actually used, the process of enriching texts with information is then taken to the intermediate step of surface syntax annotation, so-caled analytical level (Hajič 1998). A dependency tree structure is assigned to each sentence, the nodes of the tree are labelled with word tokens and the basic functions of individual words in the sentence. This level of annotation distinguishes a relatively limited set of functions (26 labels, such as predicate, subject, object, attribute, adverbial, out of which 12 are rather technical , such as sentence boundaries, labels for punctuation marks etc.). The (hitherto) final step (depth) of annotation is the tectogrammatical layer (Hajičová 1998, Hajičová and Pajas 2000, Hajičová and Pajas 2002, Böhmová 2001). On this layer, only autosemantic words get a node of their own and nodes are reestablished for elements that are deleted in the surface shape of the sentence. The overall structure is again a dependency tree; there are attributes with the total of 140 values for underlying syntactic relations, enriched by three basic values of the information structure (topic-focus articulation) and specific attributes for coreferential (inter- and intra-sentential) links.

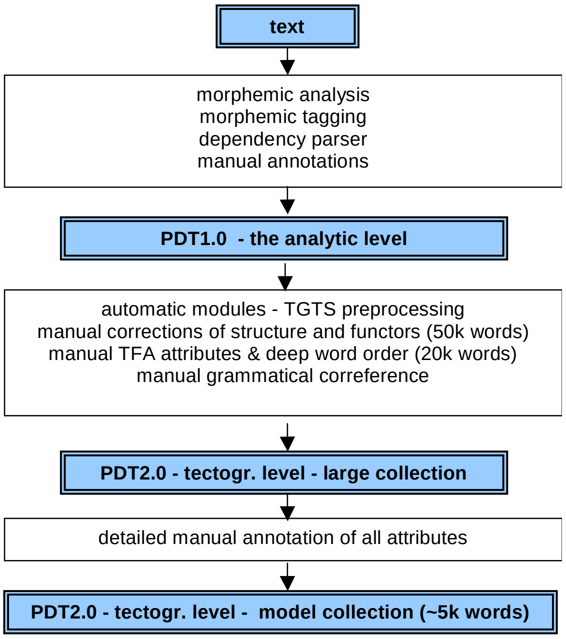

text
morphemic analysis
morphemic tagging
dependency parser
manual annotations
PDT1.0 - the analytic level
automatic modules - TGTS preprocessing
manual corrections of structure and functors (50k words)
manual TFA attributes & deep word order (20k words)
manual grammatical correference
PDT2.0 - tectogr. level - large collection
detailed manual annotation of all attributes
PDT2.0 - tectogr. level - model collection (~5k words)

Figure 1. Annotation scenario of the Prague Dependency Treebank

On the **scale of robustness** we are moving alongside the same track. It occurs that as we are getting deeper as for the level of description, the automatic procedures (both the statistical and the structural ones) provide less accurate results and the output needs more corrections done by hand.

The statistical methods used for morphemic tagging are as accurate as 95 %. This result offers morphological annotation of large amounts of texts with minimal human intervention. The whole of the first release of the Czech National Corpus (100 millions of occurrences of words) has been annotated by disambiguated tags.

On the analytical level, there are two separate statistical automatic modules involved: a parser and a procedure assigning the syntactic functions. The Collins' parser adapted to dependency grammar assigns the tree structure to the sentence with accurracy of 80 % (Collins et al. 1999), measured by the number of misplaced dependencies. Human annotators use an interactive tree editor to correct the tree structure and then they run the function assignment in the form of a macro command of the editor. The precision of this procedure reaches 92 % of accurracy. 100 thousand sentences (about 1 400 000 occurrences of words) have been annotated on this level, and published as PDT version 1.0 (see the reference below). The annotation process itself brings important material for refining some subtle points of the theory. The data are used for linguistic research, they create the basis for further annotation automation, and they are also used in research projects in the fields of information retrieval and machine translation.

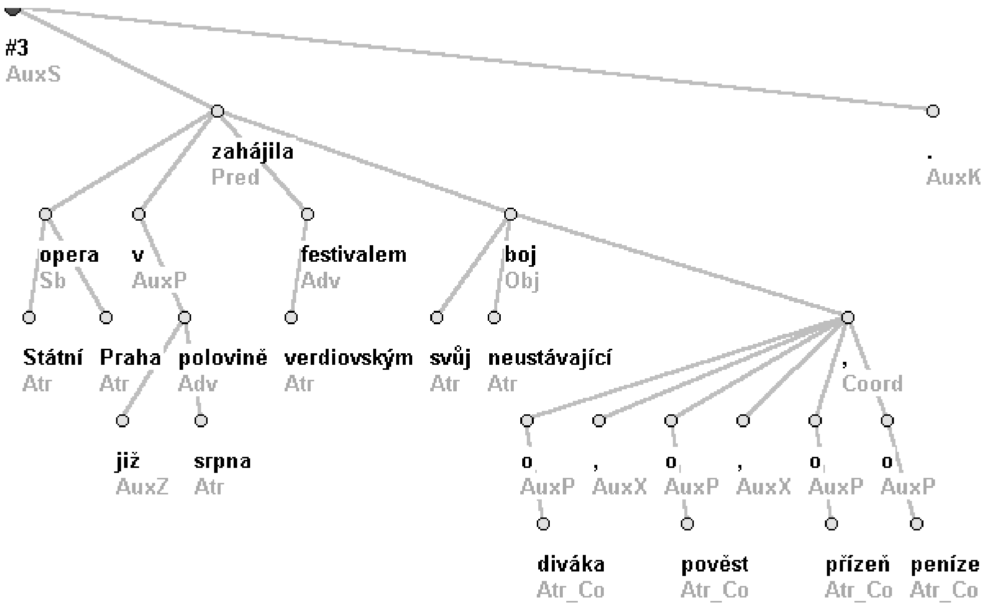


Figure 2. Analytic tree structure

The transduction from the analytical tree structures (ATS) towards the tectogrammatical tree structures (TGTS) has been divided into 3 streams of detail and 3 steps of manual procedures.

The automatic pre-processing module (Böhmová 2001) transduces the ATSs and the manual annotators complete the annotations (for a possibility to enlarge the automatic procedure as to include the assignment of functors, see Žabokrtský 2000). We are aware that we can only annotate a limited number of the sentences in full, using all the features defined in the theory, as this is a complicated resource-consuming task. Therefore we have decided to build a 'large collection' of TGTSs, annotated only as for the structure, i.e. functors are assigned for different types of dependency relations (the classification of functors is based on the FGD's theoretical approach to valency, see e.g. Panevová 1974-75; Panevová and Hajičová 1984, Sgall et al. 1986, for a similar though more cognitively oriented approach see Fillmore et al. 2002), topic-focus articulation (Hajičová 2002) and basic coreference features, and a small 'model collection' containing the complete set of features (e.g. a more detailed classification of morphological and deep syntactic functions, basic values of textual coreference).

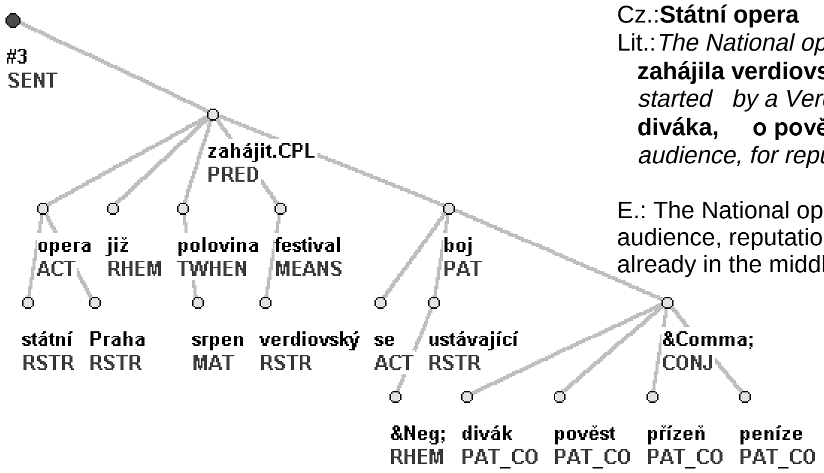
As the complete TGTS is a structure containing detailed grammatical, syntactic and coreferential information for each node, it is impossible to annotate all the features in a single step.Therefore, the annotations of the large collection are cascaded: first, the tree structure and syntactic functions are corrected (this collection contains now about 50 000 sentences), then the attributes of topic-focus articulation and deep word order are assigned (now about 10 000 sentences are fully annotated after the first and second pass), and in the last pass through the data, the values of grammatical coreference are being added.

The large collection provides data for further empirical investigations and a basis for the possible future fourth layer of annotation, namely some kind of formal semantic representation. The model collection's size is a few hundreds of sentences. The attributes specific for this collection have all been practically annotated by hand. The collection serves for further theoretical research and for a possible complementation of automatic procedures.

The combination of all automatic parts of the annotation into a complete transduction from text to tectogrammatical trees, without the manual correction between the steps, was a part of our more recent machine translation project (http://www.cslp.jhu.edu/ws2002/groups/mt/). The preliminary results show that the main share of errors is caused by coordination constructions. The automatically generated 11000 TGTSs have successfully modeled the data needed for the automatic translation.

Another view we have considered is the **scale of coverage**. We have made the decision not to restrict the set of sentences that we are processing. Prague Dependency Treebank uses real, non-adapted running texts from different sources (daily newspapers, scientific magazines, popular scientific literature etc.) and the texts have been selected to create a balanced set of texts from these sources, without limiting ourselves to certain length or structural complexity of the sentence. The annotation scheme contains instructions (more or less technical) for assigning the structure and syntactic funtions to addresses, tables, titles and other special types of sentences. A very important issue is that of the tools available for the annotators: for the purpose of PDT annotations, a highly annotator friendly, but internally sophisticated software tool has been built to process real input (see Mírovský, Ondruška 2002; Mírovský, Ondruška, Průša 2002; Hajič, Pajas, Vidová-Hladká 2001).

Figures 2 and 3 show examples of the same sentence annotated on the analytical and tectogrammatical level, respectively.



Cz.:**Státní opera Praha již v polovině srpna**
Lit.:*The National opera Prague already in the middle of August*
zahájila verdiovským festivalem svůj neustávající boj o
started by a Verdi festival its neverending quest for
diváka, o pověst, o přízeň, o peníze.
audience, for reputation, for support, for money.

E.: The National opera Prague started its neverending quest for audience, reputation, support and money by a Verdi festival already in the middle of August.

Figure 3. Tectogrammatical tree structure

3. Conclusion

We believe that the multi-layer scenario of PDT together with the division of labour into three annotation streams and four levels of depth of description, as well as the combination of automatic, semi-automatic and hand-crafted procedures used, serve well for the aim to balance the accuracy of the annotation and the requirements imposed by the amount of annotated data.

Acknowledgements

The work on the project reported on in this paper is supported by a grant of the Czech Ministry of Education, LN00A063.

Bibliography

Böhmová, Alena (2001). *Automatic Procedures in Tectogrammatical Tagging*. The Prague Bulletin of Mathematical Linguistics 76, MFF UK Prague, pp 23 - 34

Böhmová, Alena; Hajič, Jan; Hajičová, Eva and Barbora Vidová-Hladká (2003*): The Prague Dependency Treebank: A Three-Level Annotation Scenario*. In: Treebanks: Building and Using Parsed Corpora (ed. Anne Abeille) Kluwer Academic Publishers

Collins, Michael; Hajič, Jan; Ramshaw, Lance and Christoph Tillmann(1999) *A Statistical Parser for Czech*. In: Proceedings of 37th ACL'99 Maryland, USA, pp 22 - 25

Dipper S., Brants T., Lezius W., Plaehn O. and G. Smith (2001) *The TIGER Treebank*. Paper presented at the LINC-2001 workshop at the Annual Meeting of SLE, Leuven

Erk K., Kowalski A., Padó S. and Pinkal M. (2003). *Towards a resurce for lexical semantics; A large German corpus with extensive semantic annotation*. In Proceedings of the 41st Annual meeting of the ACL, Sapporo, Japan, pp 537-544.

Fillmore C.J., Baker C.F. and H. Sato (2002). *The FrameNet database and software tools*. In Proceedings of 3rd International Conference on Language Resources and Evaluation, vol IV, Las Palmas.

Hajič, Jan (1998). *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*. In: Issues of Valency and Meaning. (ed. Eva Hajičová) Karolinum, Charles University Press, Prague, pp 106 - 132

Hajič Jan (2003). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Charles University Press - Karolinum

Hajič, Jan; Hajičová, Eva; Holub, Martin; Řezníčková, Veronika; Pajas, Petr; Sgall, Petr and Barbora Vidová-Hladká (2001). *The Current Status of the Prague Dependency Treebank*. TSD2001 Proceedings (eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer), LNAI 2166. Springer-Verlag Berlin Heidelberg New York, ISBN 3-540-42557-8, pp. 11-20

Hajič, Jan; Pajas, Petr; Vidová-Hladká, Barbora (2001*). The Prague Dependency Treebank: Annotation Structure and Support*. Proceedings of the IRCS Workshop on Linguistic Databases. Univesity of Pennsylvania, Philadelphia, USA, pp. 105-114

Hajičová, Eva (1998). *Prague Dependency Treebank: From analytic to tectogrammatical annotations*. Text, Speech, Dialogue. Proceedings of the First Workshop on Text, Speech, Dialogue-TSD`98 (eds. P. Sojka, V. Matou□ek, K. Pala, I. Kopeček). Brno: Masaryk University, Czech Republic, ISBN 80-210-1900-X

Hajičová, Eva (1999). *The Prague Dependency Treebank: Crossing the Sentence Boundary*. TSD'99, Proceedings (eds. V. Matoušek, P. Mautner, J. Ocelíková, P. Sojka). Lecture Notes in Artificial Intelligence vol.1692, Springer

Hajičová, Eva (2002). *Theoretical description of language as a basis of corpus annotation: The case of Prague Dependency Treebank*. Prague Linguistic Circle Papers 4, pp. 111-127. Amsterdam/Philadelphia: Benjamins

Hajičová, Eva (2002). *Topic-focus articulation in the Czech National Corpus*. In: Josef Hladký, ed., Language and Function. To the Memory of Jan Firbas. Amsterdam/Philadelphia: John Benjamins Publ. Company, pp. 185-194

Hajičová, Eva and Petr Pajas (2000). *Evaluation of Tectogrammatical Annotation of PDT*. TSD2000, Proceedings (eds. P. Sojka, I. Kopeček, K. Pala), pp. 75-80. Lecture Notes in Artificial Intelligence vol.1902, Springer, ISBN Lecture Notes in Artificial Intelligence vol.1902, ISBN 3-540-41042-2 , 2000

Hajičová, Eva and Petr Pajas (2002). *Corpus annotation on the tectogrammatical layer: Summarizing of first stages of evaluation*. PBML 77, pp. 5-18. MFF UK, Prague

Kingsbury P., Palmer M. and M. Marcus (2002). *Adding semantic annotation to the Penn Treebank.* In Proceedings of the Human Language Technology Conference, San Diego, California.

Mírovský, Jiří; Ondruška, Roman (2002). *Netgraph System-Searching through Prague Dependency Treebank*. Prague Bulletin of Mathematical Linguistics 77

Mírovský, Jiří; Ondruška, Roman; Průša, Daniel (2002*). Searching through Prague Dependency Treebank-Conception and Architecture*. Proceedings of "The First Workshop on Treebanks and Linguistic Theories", 20th and 21st September 2002, Sozopol, Bulgaria, pp. 114-122. LML, Bulgarian Academy of Sciences, Sofia, Bulgaria and SfS, Tuebingen University, Tuebingen, Germany

Hajičová Eva; Panevová Jarmila and Petr Sgall (2000). *A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank.* UFAL/CKL Technical Report TR-2000-09, Charles University, Prague, Czech Republic

Skut W., Krenn B., Brants T. and Uszkoreit H. (1998*). A linguistically interpreted corpus of German newspaper text*. In Proceedings of LREC 98, Granada, Spain

Vidová-Hladká, Barbora (2000). *The Context (not only) for Human*. LREC (2nd Intern. Conference), vol.II (eds. M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer), pp. 1113-1116. Athens, Greece

Žabokrtský, Zdeněk (2000). *Automatic Functor Assignment in the Prague Dependency Treebank*. TSD2000, Proceedings (eds. P. Sojka, I. Kopeček, K. Pala), pp. 45-50. Lecture Notes in Artificial Intelligence vol.1902, Springer, ISBN 3-540-41042-2

Generation in the context of MT, CLSP workshop 2002, Baltimore, http://www.cslp.jhu.edu/ws2002/groups/mt/

PDT online resources: http://ufal.mff.cuni.cz/pdt/