Utilizing Linguistic Resources
Theory and Practical Experience

Vašek Němčík

NLP Laboratory
Faculty of Informatics, Masaryk University
Brno, Czech Republic
xnemcik@fi.muni.cz

Abstract. The Prague Dependency Treebank (henceforth PDT) is a large collection of texts in Czech. It contains several layers of rich annotation, ranging from morphology to deep syntax. It is unique in its size and theoretical background, especially for a language like Czech, which can be, with regard to the number of its speakers, considered a small language. In this article, we use PDT 2.0 to demonstrate that within real NLP systems, complex annotations may cut both ways. We present several issues that might pose problems when extracting data from PDT, and complex structures in general, and hint on possible solutions.

1 Introduction

The Prague Dependency Treebank 2.0 (henceforth just PDT) is a large collection of Czech texts compiled at the Institute of Formal and Applied Linguistics at the Charles University in Prague. It is probably the most notable linguistic resource available for the Czech language. Taking into account that Czech is a rather "small" language by the number of its speakers, PDT can be considered unique as it exhibits a very flattering combination of large corpus size and annotation richness. The main aim of the work on PDT was to yield a resource that would allow testing the theoretical claims following from the long tradition of the Functional Generative Description of language by confronting them with real data. Further motivation was to obtain data for training machine-learning based NLP applications.

Without any doubt, the emergence of PDT has made it possible to study linguistic phenomena that were not easy to investigate on a large scale before. On the other hand, with all respect to the long-standing tradition of FGD and the work done on PDT, it can be argued that the potential of PDT as training corpus for real NLP systems could be further extended by simplifying the data structures.

In the next section, we overview The Prague Dependency Treebank and its main features. Next, in Section 3, we mention selected features of the PDT annotation that might be considered rather unfortunate for practical purposes, as illustrated on a particular NLP task. Finally, we conclude the paper by reviewing the presented ideas and sketching the plans of further work.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 47–51, 2010.

2 The Prague Dependency Treebank

The PDT is an open-ended project for manual annotation of substantial amount of Czech-language data with linguistically rich information ranging from morphology through syntax and semantics/pragmatics and beyond [1]. Its central point is a large corpus containing the mentioned annotation, now available in version 2.0. It is probably the most notable linguistic resource available for Czech. Taking into account its size and richness of annotation, and a rather limited number of speakers of Czech, it can be even cosidered unique.

The information about texts in PDT is organized as multi-layer annotation. The overview of layers available is presented in Table 1.

Table 1. The layers in PDT 2.0

annotation layer	*brief characterization*	*size*
morphological layer	(words and morphological features)	2 million words
analytical layer	(syntactic dependency trees)	1.5 million words
tectogrammatical layer	(trees with deep sentence structure)	0.8 million words

The first layer of annotation, **the morphological layer**, contains information about grammatical features of individual words. For each sentence in the corpus, it contains a linear list of words, accompanied by their respective morphological tags. More information about the features and their possible values within this layer can be found in [2].

The analytical layer comprises of the same tokens as the previous layer, however, their organization is not linear. On this layer, they form a tree based on the relation of syntactic dependency.

The nodes of the trees (i.e. the individual words) contain information about further features. The most relevant is probably the analytical function describing the grammatical role of the word (subtree) in the sentence. Each node also carries information about its linear position in the sentence, and also a link to the related token on the morphological layer. The linking between the a-layer and m-layer is one-to-one, i.e. each node on the a-layer has a corresponding token on the m-layer, and vice versa. More information about the analytical trees can be found in [3].

The highest level level of PDT 2.0, **the tectogrammatical layer**, captures diverse linguistic aspects beyond syntax. Each sentence is represented by a dependency tree reflecting the deep structure of each sentence.

The nodes of the tectogrammatical tree carry various further information. Each node carries its semantic role with regard to its structural mother, and where applicable, nodes carry information about its valency. Further, tectogrammatic trees contain information about grammatemes of auto-semantic words, topic-focus articulation, coreference, etc. Notably, the linking between

the t-layer and a-layer is *not* one-to-one. Many nodes of the analytical layer have been ommitted (e.g. prepositions and punctuation), on the other hand, new nodes have been added (e.g. representations of people or things that are semantically present, but not explicitly voiced in the sentence). More information on features stored in tectogrammatical trees is revealed in [4].

The annotation was carried out manually, based on outputs of various automatic tools that yield an approximate form of the relevant information.

3 Practical Issues

As mentioned above, PDT is a very enticing linguistic resource, both in its size and the scope of phenomena it encompasses. On the top of that, it is supported by the underlying Functional Generative Description theory, which has a long and respectable tradition in general linguistics, and there is hardly any doubt about its consistency.

On the other hand, designing successful NLP systems often requires a rather different, practical, and sometimes even slightly heretic, approach to language. It is not crucial that the system is based on a sophisticated linguistic theory, and that it handles marginal phenomena correctly, as long as it performs reasonably with regard to its purpose. This concerns both the algorithm design, and the data used within the system. Usually it is of great advantage when the underlying principles are rather straightforward.

This is obviously the case with the notion of syntactic dependency which is central in the Praguian linguistic tradition. The notion of one word being syntactically dependent on some other word, is computationally very feasible, and also the theoretical consequences are very straightforward.

The dependency trees within the PDT, however, are not plain dependency trees. Apart from dependency edges, they also contain edges of various other types. These edges mainly account for coordination and apposition. At first glance, this seems to be a very clean and elegant solution, however, together with the convention of attaching arguments of coordinated nodes to the respective conjuction, it alters the tree structure considerably. Most importantly, it has a rather unfortunate consequence, namely that unlike in a plain dependency tree, a phrase is not necessarily a (sub)tree. This fact makes processing of the tree data rather cumbersome.

This can be demonstrated on a sample (yet very real) processing task – detection of unvoiced subjects of clause predicates. PDT seems to be a suitable source of training or evaluation data for this task. However, extracting this type of data from PDT is not as straightforward as it may seem.

Firstly, PDT does not explicitly contain information about clauses. This seems to be a consequence of the fact that the notion of clauses is somewhat irrelevant from the dependency point of view. Unfortunately, for many NLP tasks, such as re-construction of missing subjects in pro-drop languages, it is the main processing unit.

The next step and a logical way towards our goal would be to detect clauses and their predicates based on the information stored on the analytical layer. This can be done procedurally, by traversing a-layer trees in a top-down manner. However, this process is rather cumbersome. Verbal nodes representing a clause predicate are not easy to distinguish from infinitives as the relevant auxiliary verbs, modal verbs, and other relevant nodes might possibly be at various positions of the tree. So might be the node representing the subject, the presence (or absence) of which is the key point of our investigation. As a result of this, we arrive at a heuristic procedure detecting clause boundaries and missing subjects, with a non-zero error rate.

This is rather disappointing, as the information we need to extract from PDT seems to be a key factor for various decisions during the annotation process, both on the analytical, and the tectogrammatical layer. In practice, a comparably feasible alternative to extracting this information from PDT would probably be computing this information from plain text using shallow parsing and simple heuristics.

The use of information on the tectogrammatical layer in real-life NLP systems lies probably in the future as most of the data can't be obtained automatically with a satisfactory reliability by contemporary systems. However, a considerable obstacle in practical usability of t-layer trees seems to be their rather complex structure. Apart from the constructions common on the analytical layer, there are further phenomena that have an impact on the basic notion of dependency in the t-layer trees, such as several types of newly generated trees and linkings. Studying the representational conventions of the tectogrammatical layer to prevent unexpected results, is a rather time-consuming task as the available annotation manual consists of 1215 pages. Unfortunately, this fact as such means a significant motivation to search for alternative data sources. It also inevitably raises the question whether it is possible for a human, as error-prone as they are in their essence, to produce consistent and reliable annotation based on such large and complex annotation guidelines. These psychological effects are rather unfortunate as these doubts are probably hollow.

This is an interesting contrast to projects such as The Sketch Engine, which is based on simple, however, from the linguistic point of view not particularly clean ideas. The contrast suggests that also in the world of language technology, simplicity is at least as appealing as a wide range of features.

4 Conclusions and Further Work

This paper reviewed the main features of The Prague Dependency Treebank, and its annotation levels. Further it described certain difficulties that may arise when using complex linguistic data in a practical NLP setting.

The presented obstacles in extracting a specific type of information from PDT hints that richness of data structures is not always a clear advantage. A stricter (simpler) implementation of the dependency principle within the tree structures might make data easier to use. As our future work, we plan to refine

our heuristics for extracting clauses and unvoiced clause subjects from the PDT annotation and to export it into a simple, linear token-based format.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Hajič, J., et al.: The Prague Dependency Treebank 2.0. Developed at the Institute of Formal and Applied Linguistics, Charles University in Prague. (2005) `http://ufal.mff.cuni.cz/pdt2.0/`.
2. Zeman, D., Hana, J., Hanová, H., Hajič, J., Hladká, B., Jeřábek, E.: A Manual for Morphological Annotation, 2nd edition. Technical Report 27, ÚFAL MFF UK, Prague, Czech Republic (2005).
3. Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A.: Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory. Technical Report 28, ÚFAL MFF UK, Prague, Czech Republic (1999).
4. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová-Řezníčková, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z.: Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory. Technical report, ÚFAL MFF UK, Prague, Czech Republic (2005).