Annotation Tool for Discourse in PDT

Jiří Mírovský, Lucie Mladová, Zdeněk Žabokrtský
Charles University in Prague
Institute of Formal and applied Linguistics
{mirovsky,mladova,zabokrtsky}@ufal.mff.cuni.cz

Abstract

We present a tool for annotation of semantic inter-sentential discourse relations on the tectogrammatical layer of the Prague Dependency Treebank (PDT). We present the way of helping the annotators by several useful features implemented in the annotation tool, such as a possibility to combine surface and deep syntactic representation of sentences during the annotation, a possibility to define, display and connect arbitrary groups of nodes, a clause-based compact depiction of trees, etc. For studying differences among parallel annotations, the tool offers a simultaneous depiction of parallel annotations of the data.

1 Introduction

The Prague Dependency Treebank 2.0 (PDT 2.0; Hajič et al., 2006) is a manually annotated corpus of Czech. It belongs to the most complex end elaborate linguistically annotated treebanks in the world. The texts are annotated on three layers of language description: morphological, analytical (which expresses the surface syntactic structure), and tectogrammatical (which expresses the deep syntactic structure). On the tectogrammatical layer, the data consist of almost 50 thousand sentences.

For the future release of PDT, many additional features are planned, coming as results of several projects. Annotation of semantic inter-sentential discourse relations (Mladová et al., 2009)[1] is one of the planned additions. The goal is not only to annotate the data, but also to compare the representation of these relations in the Prague Dependency Treebank with the annotation done at the Penn Treebank, which was carried out at University of Pennsylvania (Prasad et al., 2008).

Manual annotation of data is an expensive and time consuming task. A sophisticated annotation tool can substantially increase the efficiency of the annotations and ensure a higher inter-annotator agreement. We present such a tool.

2 Tree Editor TrEd and the Annotation Extension

The primary format of PDT 2.0 is called PML. It is an abstract XML-based format designed for annotation of linguistic corpora, and especially treebanks. Data in the PML format can be browsed and edited in TrEd, a fully customizable tree editor (Pajas and Štěpánek, 2008).

TrEd is completely written in Perl and can be easily customized to a desired purpose by extensions that are included into the system as modules. In this paper, we describe the main features of an extension that has been implemented for our purposes. The data scheme used in PDT 2.0 has been enriched too, to support the annotation of the discourse relations.

2.1 Features of the Annotation Tool

A tool for the annotation of discourse needs to offer several features:

- creation of a link between arguments of a relation
- exact specification of the arguments of the relation

[1] It is performed in the project *From the structure of a sentence to textual relations* (GA405/09/0729), as one of several tasks.

- assigning a connective to the relation
- adding additional information to the relation (a type, a source, a comment etc.)

Links between arguments: The annotation of discourse relations in PDT is performed on top of the tectogrammatical (deep syntactic) layer of the treebank. Similarly to another extension of TrEd, dedicated to the annotation of the textual coreference and the bridging anaphora (Mírovský et al., 2010), a discourse relation between nodes is represented by a dedicated attribute at the initial node of the relation, containing a unique identifier of the target node of the relation.[2] Each relation has two arguments and is oriented – one of the arguments is initial, the other one is a target of the link. The link is depicted as a curved arrow between the nodes, see Figure 1. Although the arrow connects the two nodes, it does not mean that the two nodes themselves equal the two arguments of the relation – more about it later.

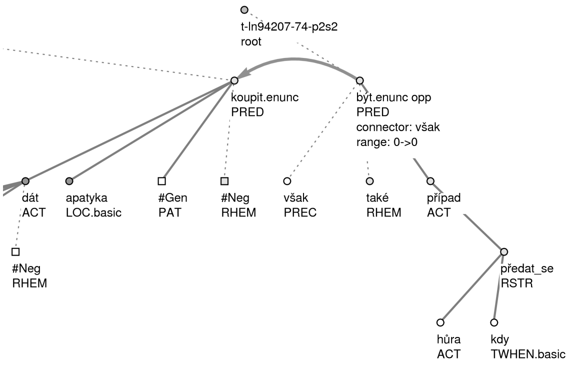

Figure 1. An arrow represents a link.

Additional information about the relation is also kept at the initial node – there is an attribute for the type, an attribute for the source (who annotated it) and an attribute for a comment.

Extent of the arguments: Usually, an argument of a discourse relation corresponds to a subtree of a tectogrammatical tree and can be represented simply by the root node of the subtree. However, there are exceptions to this "rule". Sometimes it is necessary to exclude a part of the subtree of a node from the argument, sometimes the argument consists of more than one tree and sometimes it is even impossible to set exactly the borders of the argument. To allow for all these variants, each discourse link has two additional attributes specifying range of the initial/target argument (both are stored at the initial node of the link). The possible values are:

- "0" (zero) – the argument corresponds to the subtree of the node
- *N* (a positive integer) – the argument consists of the subtree of the node and of *N* subsequent (whole) trees
- "group" – the argument consists of an arbitrary set of nodes (details below); this should only be used if the previous options are not applicable
- "forward" – the argument consists of the subtree of the node and an unspecified number of subsequent trees; should only be used if more specific options are not applicable
- "backward" – similarly, the argument consists of the subtree of the node and an unspecified number of preceding trees; should only be used if more specific options are not applicable

Groups: An argument of a discourse relation can consist of an arbitrary group of nodes, even from several trees. The fact is indicated in a range attribute of the relation (by value "group"). Another attribute then tells which group it is. Groups of nodes inside one document are identified by numbers (positive integers). Each node can be a member of several groups; a list of identifiers of groups a node belongs to is kept at the node. Every group has a representative node – if a discourse link starts/ends at a group, graphically it starts/ends at the representative node of the group, which is the depth-first node of the group belonging to the leftmost tree of the group. Figure 2 shows an example of a group. In the example, the right son (along with its subtree) of the target node of the relation has been excluded from the target argument of the relation (by specifying the target group of nodes, which is graphically highlighted). The right son (and its subtree) is actually the initial argument of the relation.

[2] The data representation allows for several discourse links starting at a single node – there is a list of structured discourse elements representing the individual relations.

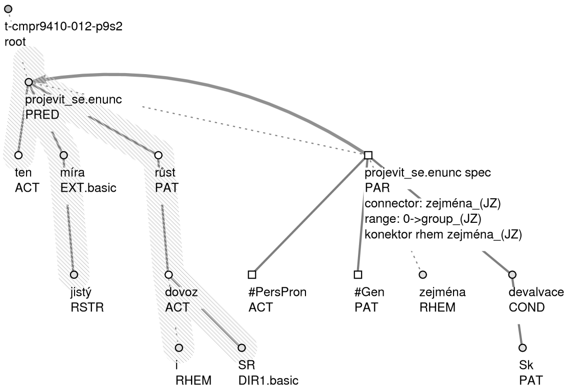

Figure 2. A group of nodes.

Connectives: A connective of a discourse relation is represented as a list of identifiers of (usually) tectogrammatical nodes that correspond to the surface tokens of the connective; the list is kept at the initial node of the relation. It is often only one node, sometimes it consists of several nodes. However, some tokens (like a colon – ":") are not represented on the tectogrammatical layer (at least not as a node). Therefore, identifiers of nodes from the analytical layer are allowed as well.

Collapsed trees: To be able to display more information using less space, a collapsed mode of depicting trees has been implemented.

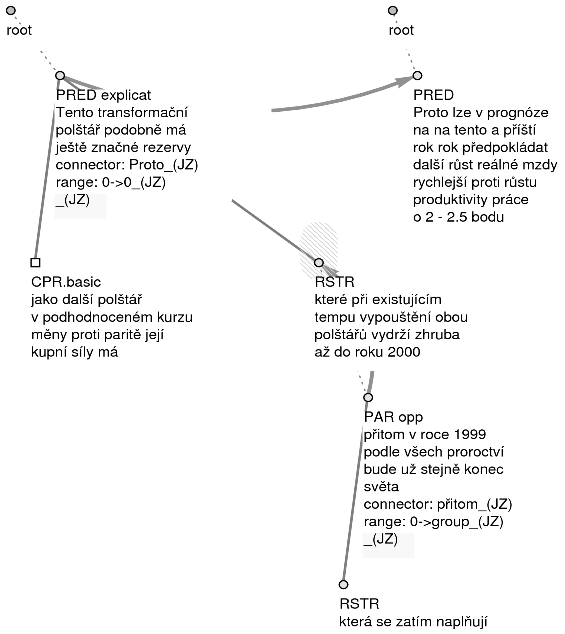

Figure 3. A collapsed mode of depicting trees.

A simple algorithm based on the tectogrammatical annotation has been employed to collapse each subtree representing an individual clause of the sentence into one node. Figure 3 shows an example of two collapsed trees.

Discourse relations most often start/end at nodes representing roots of the clauses. In those rare cases when the discourse relation should lead inside a clause, the annotators can un-collapse the trees, create the link, and collapse back. Such a link would then be depicted with a dotted arrow.

Other features: The tool also incorporates some other features that make the annotation of discourse relations easier. Based on their preference, the annotators can annotate the relations either on the trees or on the linear form of the sentences in the text window of the tool. In the sentences, the tokens that represent the initial/target nodes of the relations are highlighted and easily visible.

2.2 Parallel Annotations

To study discrepancies in parallel annotations, a mode for depicting parallel annotations exists. It can display annotations of the same data from two or more annotators. Figure 4 shows parallel annotations from two annotators. In this example, the two annotators ("JZ" and "PJ") agreed on the relation on the top of the figure, they also marked the same connective ("Poté"), and selected the same type of the relation ("preced(-ence)"). They also agreed on the range of both the arguments ("0", i.e. the subtrees of the nodes). The other relation (on the left, below the first one) has only been recognized by one annotator ("JZ").

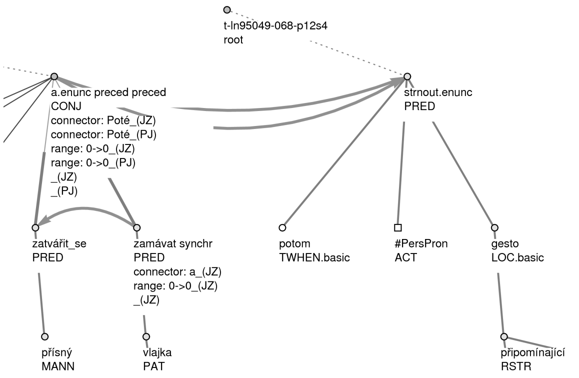

Figure 4. Parallel annotations.

3 Conclusion

From the technical point of view, we have described features of an annotation tool for semantic inter-sentential discourse relations in the Prague Dependency Treebank 2.0. We have shown how it (hopefully in a simple and intuitive manner) allows for quite complex configurations of arguments, and offers features that make the annotation easier. A mode for studying parallel annotations has also been implemented.

Evaluation of such a tool designed for a highly specific task is difficult, as the tool does not produce any direct results (apart from the annotated data) and is highly adapted to our – given the tectogrammatical trees – quite unique needs. (The annotated data themselves, of course, can be (and have been, see Zikánová et al., 2010) evaluated in various ways.) Bird and Liberman (2001) listed some very general requirements on annotation tools for linguistic corpora, namely:

- generality, specificity, simplicity,
- searchability, browsability,
- maintainability and durability.

The first requirement applies both to the annotation tool and the annotation framework. As described e.g. in Mladová et al. (2009), the annotation framework that we use is based on the knowledge obtained from studying various other systems, especially the Penn Discourse Treebank (Prasad et al., 2008), but naturally it has been adjusted to specific needs of the Czech language and PDT. The inter-connection of our system with the tectogrammatical layer of PDT helps in some annotation decisions, as many ambiguities have already been solved in the tectogrammatical annotation.

The second requirement – searchability and browsability – is very easily fulfilled in our framework. A very powerful extension for searching in PML-formatted data, called PML Tree Query, is available in TrEd (Pajas and Štěpánek, 2009).

PML is a well defined formalism that has been used extensively for large variations of data annotation. It can be processed automatically using btred, a command-line tool for applying Perl scripts to PML data, as well as interactively using TrEd. Therefore, we believe that our annotation framework and the annotation tool fulfill also the third requirement.

Acknowledgments

We gratefully acknowledge support from the Czech Ministry of Education (grant MSM-0021620838), and the Grant Agency of the Czech Republic (grants 405/09/0729 and P406/2010/0875).

References

Bird S. and M. Liberman. 2001. *A formal framework for linguistic annotation.* Speech Communication 33, pp. 23–60.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and M. Ševčíková-Razímová. 2006. *Prague Dependency Treebank 2.0.* CD-ROM, LDC2006T01, Linguistic Data Consortium, Philadelphia, USA.

Mladová, L., Zikánová, Š., Bedřichová, Z., and E. Hajičová. 2009. *Towards a Discourse Corpus of Czech.* Proceedings of the fifth Corpus Linguistics Conference, Liverpool, UK.

Mírovský, J., Pajas, P., and A. Nedoluzhko. 2010. *Annotation Tool for Extended Textual Coreference and Bridging Anaphora.* Proceedings of LREC 2010, European Language Resources Association, Valletta, Malta.

Pajas, P. and J. Štěpánek. 2008. *Recent advances in a feature-rich framework for treebank annotation.* Proceedings of Coling 2008. Manchester, pp. 673–680.

Pajas, P. and J. Štěpánek. 2009. *System for Querying Syntactically Annotated Corpora.* Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Association for Computational Linguistics, Suntec, Singapore, pp. 33–36.

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., and B. Webber. 2008. *The Penn Discourse Treebank 2.0.* Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech.

Zikánová, Š., Mladová, L., Mírovský, J., and P. Jínová. 2010. *Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank.* Proceedings of LREC 2010, European Language Resources Association, Valletta, Malta.