Annotation Tool for Extended Textual Coreference and Bridging Anaphora

Jiří Mírovský, Petr Pajas, Anna Nedoluzhko
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
E-mail: {mirovsky, nedoluzko, pajas}@ufal.mff.cuni.cz

Abstract

We present an annotation tool for the extended textual coreference and the bridging anaphora in the Prague Dependency Treebank 2.0 (PDT 2.0). After we very briefly describe the annotation scheme, we focus on details of the annotation process from the technical point of view. 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, an automatic maintaining of the coreferential chain, underlining candidates for antecedents, etc. For studying differences among parallel annotations, the tool offers a simultaneous depiction of several annotations of the same data. The annotation tool can be used for other corpora too, as long as they have been transformed to the PML format. We present modifications of the tool for working with the coreference relations on other layers of language description, namely on the analytical layer and the morphological layer of PDT.

1. Introduction

The Prague Dependency Treebank 2.0 (Jan Hajič et al., 2006) is a manually annotated corpus of Czech. The texts are annotated on three layers – morphological, analytical and tectogrammatical.

The most abstract (tectogrammatical) layer includes among other mark-ups the annotation of coreferential links. The whole corpus contains almost 50 thousand sentences.

In PDT 2.0, two types of coreference have been (mainly manually) annotated: the grammatical coreference and the textual coreference. The grammatical coreference typically occurs within a single sentence, the antecedent being able to be derived on the basis of grammatical rules of the given language. The textual coreference has been restricted up to now to cases in which a demonstrative *this* or an anaphoric pronoun of the 3rd person, also in its zero form, are used (Kučová and Hajičová, 2004).

Our paper and the demo focus on the annotation tool for the next stage of the anaphoric annotation, which is being carried out on PDT 2.0 now. In this stage, the textual coreference is annotated also for non-pronominal and non-zero NPs, and also for some cases of adjectives, adverbs and verbs. Together with this textual coreference, bridging relations of several types are being annotated (Nedoluzhko et al., 2009a).

Manual annotation of textual coreference and bridging anaphora is a costly process and requires a constant and strong focus from the annotators. The inter-annotator agreement is not particularly high (0.6 – 0.8 F-measure in PDT, see Nedoluzhko et al., 2009b) and any help the annotation tool can provide is useful.

In *section 2* we give a short overview of the annotation scheme of the ongoing project of annotating the extended textual coreference and the bridging anaphora in PDT 2.0.

In *section 3* we present the tree editor TrEd, which is a base for our annotation tool, and we describe in detail our extension of TrEd, dedicated to purposes of the annotation project. We elaborate on its features that help the annotation.

In *section 4* we present a feature of the annotation tool that offers a comprehensive depiction of several annotations of the same data (e.g. for comparing annotations of several annotators).

In *section 5* we describe a modified version of the extension, developed for purposes of the project PlayCoref. This modification allows projecting and displaying the coreference relations on surface layers of annotation – the morphological layer and the analytical layer.

We conclude in *section 6*.

2. The Extended Textual Coreference and the Bridging Anaphora in PDT 2.0

In the ongoing annotation project, we annotate two types of anaphoric expressions – the extended textual coreference for the relation of identity and the bridging anaphora for the relations between non-coreferential entities.

The textual coreference is further classified into two types – coreference of NPs with specific (type 0 (zero)) or generic (type NR) coreference. For the bridging anaphora, the following types are distinguished: PART, SUBSET and FUNCT (traditional relations), CONTRAST for coherence relevant discourse opposites, ANAF for explicitly anaphoric relations without coreference, and the further underspecified group REST. The types PART, SUBSET and FUNCT are further specified according to the linear order of the antecedent and the anaphor in the text, e.g. PART_WHOLE is used for the case when the antecedent of the anaphoric NP corresponds to the whole of which the anaphor is a part, and WHOLE_PART for the opposite.

Annotation of the textual coreference is based on the chain principle, the anaphoric entity always referring to the last preceding coreferential antecedent. For the bridging anaphora, the chain principle is not preserved. To develop maximally consistent annotation scheme, we follow a number of basic principles, such as the principle of maximal length of coreferential chains, the principle of maximal size of an anaphoric expression (subject to annotation is always the whole subtree of the antecedent/anaphor), the principle of cooperation with the syntactic structure of a given dependency tree, which does not let annotate relations that are already caught up by the syntactic structure of the tectogrammatical tree, etc. (Nedoluzhko et al., 2009b).

3. 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 treebanks. For editing and processing data in the PML format, a fully customizable tree editor TrEd has been implemented (Pajas & 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 section, we describe some features of an extension that has been implemented for our purposes.

The data scheme used in PDT 2.0 has been enriched to support the annotation of the extended textual coreference (which has – unlike the originally annotated textual coreference – a type) and the bridging anaphora (which has not been annotated before and also has a type). Technically, various kinds of non-dependency relations between nodes in PDT 2.0 use dedicated referring attributes that contain unique identifiers of the nodes they refer to.

3.1 Features of the Annotation Tool

The task of the annotation and also the task of the tool programming have been simplified in our case by the fact that the annotation is performed on the tectogrammatical trees. Orăsan (2003), in his presentation of the annotation tool PALinkA, named three basic tasks that a tool for annotation of coreference should offer:

- insertion of information not explicitly marked in the text (e.g. ellipsis, zero pronouns),
- marking of the elements in the text (e.g. noun phrases, utterances, sentences),
- marking the links between the elements.

Note that the tectogrammatical trees solve the first two tasks for us. Not explicitly marked information has already been restored in the trees and also the noun phrases, utterances and sentences can be represented by their root nodes (which is one of the already mentioned annotation principles). On the other hand, if we wanted to allow the annotators to insert words or nodes in the text or trees, TrEd itself (as a Tree Editor) provides this functionality.

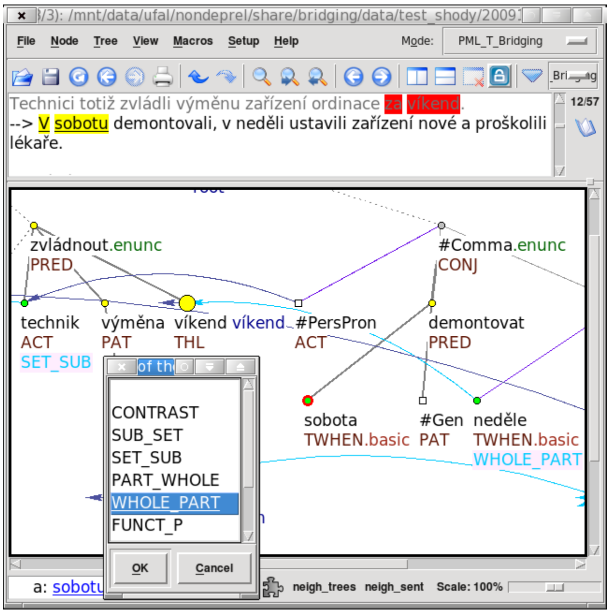
Nevertheless, the tool we have created has much more features than only the remaining third task (a possibility to mark links between elements). Various features have been implemented to help with the annotation.

Manual pre-annotation: If the annotator finds a word in the text that appears many times in the document and its occurrences seem to co-refer, he can create a coreferential chain out of these words by a single key-stroke. All nodes that have the same t_lemma (basic form of the auto-semantic word represented by the node) become a part of the chain.

Finding the nearest antecedent: The annotation instructions require that the nearest antecedent is always selected for the coreferential link. The tool automatically re-directs a newly created coreferential arrow to the nearest one (in the already existing coreferential chain) if the annotator selects a farther antecedent by mistake. However, the rule of the nearest antecedent can be broken in less clear situations. For example, if there are three coreferential words in the text, A, B and C (ordered from left to right), and the annotator connects A and C (overlooking B), and later realizes that B is also coreferential with A and creates the arrow from B to A, the tool re-connects the C→A arrow to C→B. Thus, the coreferential chain C→B→A is correctly created.

Preserving the coreferential chain: If the annotator removes an arrow and a coreferential chain is thus interrupted, the tool asks the annotator whether it should re-connect the chain.

Text highlighting: The annotation of the extended textual coreference and the bridging anaphora is performed on the tectogrammatical layer of PDT.



Technici totiž zvládli výměnu zařízení ordinace za víkend.
--> V sobotu demontovali, v neděli ustavili zařízení nové a proškolili lékaře.

Figure 1: Selecting a type of a bridging relation.

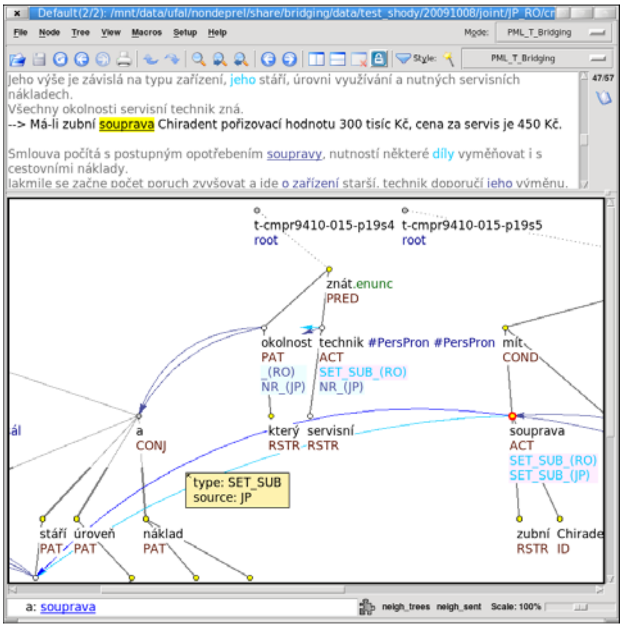
However, the annotators sometimes prefer to work on the surface form of the text, using the tectogrammatical trees only as a supporting depiction of the relations. After selecting a word in the sentences (by clicking on it), the tool determines to which node in the tectogrammatical trees the word belongs. Then, the projection back to the surface is performed and all words on the surface that belong to the selected node are highlighted. Only one word of the highlighted words is a lexical counterpart of the tectogrammatical node (which is usually the word the annotator clicked on – only in cases such as if the annotator clicks on a preposition or other auxiliary word, the lexical counterpart of the corresponding tectogrammatical node differs from the word clicked on). Using this information, also all words in the sentences that have the same t_lemma (again, we use only the lexical counterparts) as the selected word, are underlined.

Words that are connected with the selected word via a coreferential chain are highlighted in colors that indicate whether the last connecting relation in the coreferential chain is textual or grammatical. Moreover, all words that are connected via a bridging anaphora with any word of this coreferential chain, are highlighted in a specific color.

In Figure 1, the main window of the annotation tool is depicted. The annotator is in the process of creating a bridging relation between nodes “sobota” (Saturday) and “víkend” (weekend), and from a list selects the type WHOLE_PART. Darker arrows represent the textual coreference, lighter arrows represent the bridging relations.

4. Comparing Different Annotations

The tool provides a support for visual comparison of different annotations of the same data, e.g. annotations from different annotators in the inter-coder agreement measurement.


Jeho výše je závislá na typu zařízení, jeho stáří, úrovni využívání a nutných servisních nákladech.
Všechny okolnosti servisní technik zná.
--> Má-li zubní souprava Chiradent pořizovací hodnotu 300 tisíc Kč, cena za servis je 450 Kč.

Smlouva počítá s postupným opotřebením soupravy, nutností některé díly vyměňovat i s cestovními náklady.
Jakmile se začne počet poruch zvyšovat a jde o zařízení starší, technik doporučí jeho výměnu.

Default(2/2): /mnt/data/ufal/nondeprel/share/bridging/data/test_shody/20091008/joint/JP_RO/cr
File Node Tree View Macros Setup Help
Mode: PML_T_Bridging
Style: PML_T_Bridging
47/67
t-cmpr9410-015-p19s4 root
t-cmpr9410-015-p19s5 root
znát.enunc PRED
okolnost PAT _(RO) NR_(JP)
technik ACT SET_SUB_(RO) NR_(JP)
#PersPron #PersPron
mít COND
ál
a CONJ
který RSTR
servisní RSTR
souprava ACT SET_SUB_(RO) SET_SUB_(JP)
type: SET_SUB
source: JP
stáří PAT
úroveň PAT
náklad PAT
zubní RSTR
Chiradent ID
a: souprava
neigh_trees neigh_sent Scale: 100%

Figure 2: Two annotations depicted at the same time.

The data representation allows to distinguish sources of the arrows. When a file is opened in the tool, all arrows are scanned and a list of sources is created. Then, the user can choose to display arrows from all sources, or from an individual source. A pop-up window appears when the mouse pointer hovers over an arrow, and informs about the type and the source of the arrow.

Figure 2 demonstrates this feature used for the simultaneous depiction of annotations from two sources, labelled as RO and JP. The types of the arrows (displayed at the starting nodes) are also marked by the appropriate source label.

5. Re-Usability of the Tool

TrEd can and has been used not just for PDT 2.0, but also for many other data sources. The only condition is the transformation of the data to the PML format, which should be relatively easy at least for other XML-based formats using standard XML transformation techniques.

The presented extension dedicated to the coreference and bridging anaphora annotation can also be used for other corpora or similar purposes.

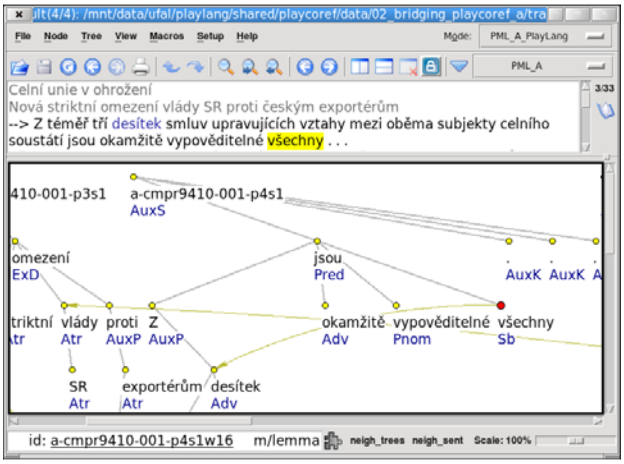
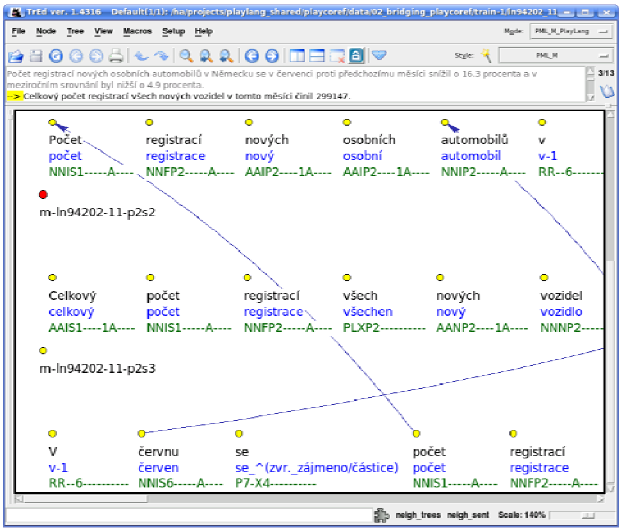

Celní unie v ohrožení
Nová striktní omezení vlády SR proti českým exportérům
--> Z téměř tří desítek smluv upravujících vztahy mezi oběma subjekty celního soustátí jsou okamžitě vypověditelné všechny . . .

Figure 3: Coreference on the analytical layer.

A modification of the extension has been created for the purposes of the project PlayCoref, whose aim is to create an on-line language game that would produce data annotated with coreference. For this project, the coreference links need to be accessed on lower layers of annotation, namely on the analytical layer and on the morphological layer. Two figures show the modified version of the extension. In Figure 3, the coreference is represented in the analytical trees. In Figure 4, coreference links are projected to the surface representation of the sentences, depicted along with the morphological information. As TrEd is a *Tree* Editor, it must be a little twisted to work with the morphological layer directly. It sees the words of a sentence as nodes of a flat tree and turns off displaying of the edges.



Počet registrací nových osobních automobilů v Německu se v červenci proti předchozímu měsíci snížil o 16.3 procenta a v meziročním srovnání byl nižší o 4.9 procenta.

--> Celkový počet registrací všech nových vozidel v tomto měsíci činil 299147.

Figure 4: Coreference on the morphological layer.

6. Conclusion

We have presented a tool for the annotation of the extended textual coreference and the bridging anaphora in PDT 2.0, along with basic principles designed for the annotation. Features of the annotation tool that help the annotators (and thus improve the quality of the annotation) have been described in detail. A mode of the tool that supports studies of the inter-annotator (dis-)agreement has also been described, as well as modifications of the tool for displaying the coreference links on other annotation layers.

It is naturally difficult to evaluate such a tool (and the annotation framework), as it does not produce comparable numbers and is designed specifically for our purposes. (The annotated data themselves, of course, can be (and have been, see Nedoluzhko et al. (2009b)) evaluated in various ways.) The focus on annotating on trees and the possibility to combine the annotation on the text and the trees make the tool stand apart from most other tools for coreference annotation. What we can do in the attempt of evaluation is find out whether the tool (and the framework) fits requirements listed in Bird and Liberman (2001), which are:

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

As described e.g. in Nedoluzhko (2009b), the annotation framework that we use is based on the knowledge obtained from studying various other systems, like MATE or GNOME, but of course has been adjusted to specific needs of the Czech language and PDT. The inter-connection of our system with the tectogrammatical layer of PDT makes it very simple, as many ambiguities have already been solved in the tectogrammatical annotation.

We definitely fulfil the second requirement – searchability and browsability. A very powerful extension for searching in PML-formatted data, called PML Tree Query, is available in TrEd (Pajas, 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 fulfil also the third requirement.

7. Acknowledgments

We gratefully acknowledge support from the Czech Ministry of Education (grant MSM-0021620838), the Grant Agency of the Czech Republic (grants 405/09/0729 and P406/2010/0875), and the Grant Agency of Charles University in Prague (GAUK 4383/2009).

8. References

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

Hajič, J. et al. (2006). Prague Dependency Treebank 2.0. *CD-ROM.* Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Hladká B., Mírovský J., and Schlesinger P. (2009). Play the Language: Play Coreference. In *Proceedings of ACL-IJCNLP 2009, main section.* Suntec, Singapore.

Kučová L. and Hajičová E. (2004). Coreferential Relations in the Prague Dependency Treebank. In *5th Discourse Anaphora and Anaphor Resolution Colloquium*. Ediçôes Colibri.

Nedoluzhko A., Mírovský J., and Pajas P. (2009a). The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of The Third Linguistic Annotation Workshop (The LAW III).* ACL-IJCNLP 2009, Suntec, Singapore, pp. 108—111.

Nedoluzhko A., Mírovský J., Ocelák R., and Pergler J. (2009b). Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. In *Proceedings of 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009).* Goa, India.

Orăsan C. (2003). PALinkA: A highly customisable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue.* ACL 2003.

Pajas, P. and J. Štěpánek J. (2008). Recent advances in a feature-rich framework for treebank annotation. In *The 22nd Interntional Conference on Computational Linguistics – Proceedings of the Conference.* Manchester, pp. 673—680.

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