Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank

Anna Nedoluzhko, Jiří Mírovský, Radek Ocelák, Jiří Pergler

Charles University in Prague
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
{nedoluzko, mirovsky}@ufal.mff.cuni.cz, radioc@seznam.cz, perglerj@volny.cz

Abstract. The present paper outlines the coding scheme for annotating extended nominal coreference and bridging relations in the Prague Dependency Treebank. We compare our annotation scheme to the existing ones with respect to the language to which the scheme is applied. We identify the annotation principles and demonstrate their application to the large-scale annotation of Czech texts. We further present our classification of coreferential relations and bridging relations types and discuss some problematic aspects in this area. An automatic pre-annotation and some helpful features of the annotation tool, such as maintaining coreferential chain, underlining candidates for antecedents, etc. are presented and discussed. Statistical evaluation is performed on the already annotated part of the Prague Dependency Treebank. We also present the first results of the inter-annotator agreement measurement and explain the most frequent cases of disagreement.

Keywords: corpus annotation, coreference, bridging anaphora

1 Introduction

The Prague Dependency Treebank (henceforth PDT) is a large collection of linguistically annotated data and documentation [2]. In PDT 2.0, Czech newspaper texts are annotated using a three layer annotation scenario. The most abstract (tectogrammatical) layer includes among other mark-ups the annotation of coreferential links. The whole corpus PDT 2.0 contains almost 50 thousand sentences.

In PDT 2.0, two types of coreference are (mainly manually) annotated: grammatical and 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. It includes coreference of relative pronouns, arguments of verbs of control, reflexive pronouns, reciprocity and verbal complements. As for textual coreference (which is not realized by grammatical means alone, but also on the basis of the context), it 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 [8].

The current paper focuses on the next stage of the coreference annotation, which is being carried out on PDT now. In this stage, the textual coreference is extended to non-pronominal and non-zero NPs, and also to some cases of adjectives, numerals and adverbs (see 3.1). Together with the textual coreference, bridging relations of several types are being annotated (see 3.3). Discourse deixis is annotated separately for references to non-nominal entities (see 3.1) and references to a discourse segment of more than one sentence (see 3.2). Section 4 deals with the annotation principles and preferences; in Section 5, we present the application of the annotation scheme on PDT and the first evaluation results. In Section 6, some cases of inter-annotator disagreement are discussed. The main problems are summarized in Section 7.

2 Previous work

Coreferential and bridging relations between discourse entities are of major importance for establishing and maintaining textual coherence. The ability to automatically resolve these kinds of relations is an important feature of text understanding systems. For both the training as well as the evaluation of these systems, manually annotated corpora are required. That is the reason why a large number of anaphoric annotation schemes have been presented just in the last few years. In what follows, we concentrate on three annotation schemes and compare them to our approach.

The MUC scheme [5] and its continuation ACE [4] are the best known and most widely used coreference schemes, developed primarily for the information extraction and other NLP tasks. Being applied to rather limited corpora, the MUC is the only existing coreference annotation scheme whose reliability has been systematically tested. Priority is given to preserving high inter-annotator agreement, so only identity relations for nouns, NPs and pronouns are annotated for coreference. The ACE program is limited to the recognition of seven entity types (person, location etc.), for which identical coreferential relations are annotated.

The MATE project, its extension on the GNOME and VENEX corpora [16] and the ongoing project of the ARRAU corpus [17] are the most well-known projects where also bridging relations are annotated. Based on MATE, the annotation scheme for coreference in Spanish was developed [18], but bridging relations have not been annotated large-scale there.

In PoCoS [6], two layers of coreference annotation schemes were suggested: the Core Scheme is general and reusable, while the Extended Scheme supports a wider range of specific extensions. The Core Scheme is used for annotating some cases of nominal coreference, while non-nominal coreference and bridging relations are annotated as part of the Extended Scheme.

All coreference annotation schemes described above consist of two steps. First, so called “markables” (the linguistic items between which coreference relations might hold) are (mostly automatically) marked, second, the relation itself is (mostly manually) defined. Markables are defined differently according to the given scheme.

In GNOME, all NPs are treated as markables, including predicative NPs, in MUC all nouns, NPs and pronouns, including 1st and 2nd person pronouns are markables, PoCoS has a sophisticated system of primary and secondary markables. Primary markables are all potential anaphors, they include definite NPs, pronouns and some other anaphoric elements. Secondary markables are e.g. indefinite NPs and are subject to annotation only if they serve as antecedents of a primary markable.

3 Coreference annotation scheme

3.1 Elements to be annotated

Unlike ACE, we do not restrict the annotation to a set of NEs, and annotate all referential entities, also the abstract and generic ones. So, the subject to annotation in PDT is actually some kind of pragmatic references to the actual notions.

The extended coreferential and bridging relations are to be marked between elements of the following categories: full NPs (*Prague – the capital of the Czech Republic*), anaphoric adverbs *(the capital of the Czech Republic - there*), numerals (*1999 – this year*), clauses and sentences if coreferring with NPs (*They tried to teach him to read – The attempt was not successful*). Similarly to MUC, adjectives are annotated only if they are coreferential with a named entity or a nominal head, so e.g. we annotate pairs as *German – Germany*. Coordinated NPs and appositional structures are also potential markables, in the syntactic structure of the tectogrammatical trees, their roots (conjunctions or punctuation marks) technically serve as coreferring nodes (see [9]).

Names and other named entities are all subjects to annotation. A substring of a named entity, however, is not to be annotated if it is not a named entity itself. Thus, for the sequence *The Charles University in Prague... Prague was...*, the two instances of *Prague* are to be marked coreferential; but in *Institute of Nuclear Research ... nuclear research* the two instances of NP *research* are neither to be coreferred nor to be marked as a bridging relation.

Contrary to MUC and ACE, predicate nominals are not considered to be coreferential with the subject, and neither the coreferential relation between appositional phrases is established.

3.2 Extended textual coreference

Extended textual coreference is marked between two elements that refer to the same object, notion or activity in the discourse. Each markable can only be the object of no more than one coreferential expression. Some exceptions to this rule for pronominal coreference [8] are being corrected by the annotation of the extended textual coreference.

Textual coreference is further subclassified into two types: coreference of NPs with specific reference (*coref_text*, type 0) and relations between NPs with generic reference (*coref_text*, type NR). Differently from other schemes (GNOME, ACE, etc.), we have not the feature of genericity assigned to all generic NPs. Nevertheless, we assume generic NPs to have other anaphoric properties in discourse, in addition they result in richer ambiguity and are the cause of lower inter-annotator agreement. These were the reasons to separate them into a special category of coreferential relations, thus forming separate coreferential chains of NPs with generic reference. Compare the following examples (all English examples are constructed in parallel to the corresponding original Czech ones):

Mary and John went together to Israel, but Mary [coref_text with "Mary", type 0] *had to return because of the illness*. (**1**)

A lion lives in a forest. I wrote my Ph.D. thesis about this animal [coref_text with "lion", type NR]. (**2**)

We do not distinguish between coreference pairs with the same lemmas (*Mary - Mary*) from the cases, in which the entities are synonymous, hyponymous/hyperonymous or are just different nominations of any other kind (*Germany – the state*, *Mary – she*, etc.). Using grammatical attributes of the tectogrammatical tree, this kind of information can be easily extracted automatically. Unlike [18], we do not annotate false positive links (lexically identical but non-coreferential NPs) as coreferential.

Special cases of textual coreference. Two special cases of (co)reference are being annotated in PDT.

First, the textual coreference covers the cases of endophoric **references to discourse segment of more than one sentence**, including also the cases, when the antecedent is understood by inferencing from a broader co-text. The pronominal anaphoras being already annotated in PDT 2.0, we add the links in which the anaphor is expressed by a full NP or an adverb, as in (3):

Celní unie bude sice existovat na papíře ještě dalších dvanáct měsíců, ale v praxi dostanou vzájemné vztahy punc tvrdosti mezinárodního obchodu. Poroste administrativa. Jistotu v tomto směru [segm] *dávají nejnovější kroky vlády SR.* (**3**)
(The custom union will formally function for twelve more months, but in fact the relations will be of a kind of international trade. The bureaucracy will go up. The latest steps of the Slovak government confirm this direction [segm].*)*

This kind of relation does not have (unlike [19]) explicitly marked antecedent, it just shows the fact that the given anaphoric NP corefers with some discourse antecedent of more than one sentence. We consider this decision to be provisional and we plan to complete it later.

Second, a specifically marked **link for exophora** denotes that the referent is "out" of the co-text, it is known only from the actual situation. In the same way as for segments, the new nominal and adverbial links are being added.

3.3 Bridging relations

Bridging relations [3] hold between two elements in which the second element is interpreted by an inferential process ("bridge") on the basis of the first one.

Unlike [19], bridging relations in PDT are annotated only between nominal expressions, no verbs are considered as anchors. Each node can only be an antecedent/anaphor for no more than one type of bridging relations.

Given that the marking of bridging relations is very useful for information extraction, question answering and other NLP tasks, we decided to annotate them in PDT. However, this is a very complicated and time-consuming task, which up to now did not show high enough evaluation results. Also the sets of bridging relations vary in different annotation schemes (see the rich variety of types in [3], seven types in MATE, and their reduction to three types (element, subset and poss - in GNOME and VENEX; part-of, set membership and thematic in [19], and part-of, set membership, and a converse relation in ARRAU).

In our project, we annotate two basic types that are widely agreed upon, and add four other types, which frequently occurred in the pilot annotation experiments and seem to be relatively reliably identifiable. The five subtypes of bridging relations in PDT are:

- ***part-of*** (prototypical example *room - ceiling*)

  This relation has two directions – the type "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).
- ***set subset/element of the set*** (prototypical example *participants – one of participants/some participants*)

  This relation is two-directional with the types SUB_SET and SET_SUB.

In some cases, the distinction between "part-of" and "set subset" groups is quite problematic, so that the only reason to decide for the type of a bridging relation is the countability of corresponding nouns. E.g. (4):

Revidoval text Prezidentské adresy. Poslední věta [bridging to "text", type WHOLE_PART or SET_SUB], *kterou v životě napsal, zněla ...*
(*He edited the <u>text</u> of President's address. The last <u>sentence</u>* [bridging to "text", type WHOLE_PART or SET_SUB] *was...*) **(4)**

For the time being, the instruction for such type of ambiguities is to annotate type PART only in clear cases of non-separable parts.

- ***object – individual function on this object*** (prototypical example *government – prime minister*);

This relation is two-directional with types P_FUNCT for the sequence object – function and FUNCT_P for the opposite.

- ***coherence relevant discourse opposites*** (type CONTRAST), e.g. (5):

People don't chew, it's cows [bridging to "people", type CONTRAST] *who chew*. (**5**)

The CONTRAST relation is not really bridging relation in a restricted sense, it could be rather labeled rhetorical or something like that. However, this kind of semantic dependence has a similar influence on the text cohesion as bridging ones. In addition, it supplements the similar kind of information in the topic-focus articulation annotation, where contrast topic is marked, and the currently annotated contrast on the discourse level [10].

- ***non-cospecifying explicit anaphoric relation*** where the anaphor is marked with a demonstrative, bridging type ANAF is used:

"Duha?" Kněz přiložil prst k tomu slovu [bridging to "duha", type ANAF], *aby nezapomněl, kde skončil*.
(*"Rainbow?" The priest put the finger on this word* [bridging to "duha", type ANAF]*, so that he didn't forget, where he stopped.*) (**6**)

- ***further underspecified group REST***

This type is used for capturing bridging references – potential candidates for a new group of bridging relations (e.g. *location – resident*, relations between relatives (*mother – son*, etc.), event – argument (*listening – listener*) and some other relations). The last type is not marked as a special group for its relatively rare occurrence in our corpus (as we do not mark verbs as bridging entities). If needed, this relation can be relatively easily extracted from the annotated data.

The participation on the text cohesion is considered to be important, so in ambiguous cases, those relations are annotated that are important for the text cohesion.

4 Annotation principles and preferences

In order to develop a maximally consistent annotation scheme, we follow a number of basic principles. Some of them are presented below:

Chain principle: Coreference relations in text are organized in ordered chains. The most recent mention of an entity is marked as the antecedent. This principle is checked automatically (see 5.1). The chain principle does not concern bridging relations.

Principle of the maximum length of coreferential chains. This principle, similar to the chain principle, concerns only the cases of textual coreference. It says that in case

of a multiple choice, we prefer to continue the existing coreference chain, rather than to begin a new one. To meet this principle, grammatical coreferential chains (already annotated in PDT) are being continued by textual ones, and similarly, the already annotated textual coreferential chains are continued by currently annotated non-pronominal links.

Principle of maximal size of an anaphoric expression. This principle says, that it is always the whole subtree of the antecedent/anaphor, which is the subject to the annotation.

Principle of cooperation with the syntactic structure of the given dependency tree. We do not annotate relations that are already captured by the syntactic structure of the tectogrammatical tree. So, unlike MUC, we do not annotate predication and apposition relations.

Also bridging relations are not to be annotated if the anaphor is a direct child of its antecedent in the tectogrammatical tree, and it has some of the predefined labels for the valency relations (functors), such as PAT(iens), AUTH(or), APP(urtenance), etc.. So, for example, the relation between *strop* (*ceiling*) and *místnost* (*room*) in the phrase

strop této místnosti
(the ceiling of the room) (**7**)

is not annotated, because in the tectogrammatical tree, the node *místnost* (*room*) has the functor APP, being the direct child of the node *strop* (*ceiling*).

Principle of preferring coreference to anaphora. Coreference, not anaphora, is subject to textual coreference annotation. In many cases, an anaphoric relation is also a coreferential relation (see e.g. (2)), this is however not always the case (e.g. (6)). In a Slavonic language lacking the grammatical category of definiteness, we cannot afford to choose only definite NPs for anaphoric annotation (as it is done e.g. in MATE and PoCoS), so we annotate all NPs that refer to the same entity. Non-coreferential anaphoric entities are annotated separately as a bridging relation (see 3.3).

Preference of coreference over bridging anaphora. The preference says that in case of multiple choice, we always prefer the textual coreference to a bridging relation. So having the following sequence of NPs:

(a)*Mary* – (b)*John* – (c)*children of the class* – (d)*Mary and John* (**8**)

we will annotate *Mary* in (8)d as coreferential to *Mary* in (8)a, rather than bridging_SUBSET to *chidren* in (8)c, although this relation would be closer.

5 Application and Evaluation

Coreference and bridging annotation is being performed using the TrEd (Tree Editor) annotation tool, developed at the Institute of Formal and Applied Linguistics at Charles University in Prague [14]. TrEd is a highly customizable tree editor developed primarily for PDT and using natively PDT data format called PML, which is an abstract XML-based format designed for the annotation of treebanks. TrEd can be used both for the manual and automatic processing of the data in PML and can be easily customized to a desired purpose by extensions that are included into the system as modules. Our decision to use TrEd for the annotation of the textual coreference and the bridging anaphora in PDT is based on the facts that it works flawlessly with the tectogrammatical tree structures (on which the annotation is performed), offers a huge set of macros predefined especially for the trees, and can be very easily adopted to our purposes. As such, it fits our needs much more than other existing tools, such as MMAX [11], or PALinkA [13].

The process of annotation consists of only one stage, "markables" being defined on the basis of the grammatical information already existing in PDT. This makes the scheme not properly comparable with other existing coreference schemes, but for the present moment, this is the best useful solution.

The annotation is carried out on tectogrammatical tree structures assigned to the sentences in the text. The present annotation scheme of PDT provides a number of coreferential attributes. The attribute *coref_gram* is used for the grammatical coreference, it contains the identifier of the antecedent. The attributes *coref_text* and *bridging* are complex attributes. They contain a structure consisting of the identifier of the antecedent and the linguistic type of the relation (see 3.2 and 3.3). If there are more than a single antecedent of one anaphor, the attribute *bridging* contains a list of these structures. The attribute *coref_special* is used for cases of coreference between a node and an entity that has no corresponding counterpart in the tectogrammatical structure: for the time being, there are two possible values of this attribute, namely *segm* in case of a coreferential link to a whole segment of the preceding text, and *exoph* in case of an exophoric relation (see 3.2.1).

Coreference relations are depicted by arrows leading from the anaphor to the antecedent and the types of relations (bridging, textual, grammatical) are distinguished by different colors of the arrows. In Figure 1, we present an example of coreference assignment by means of links used by annotators.

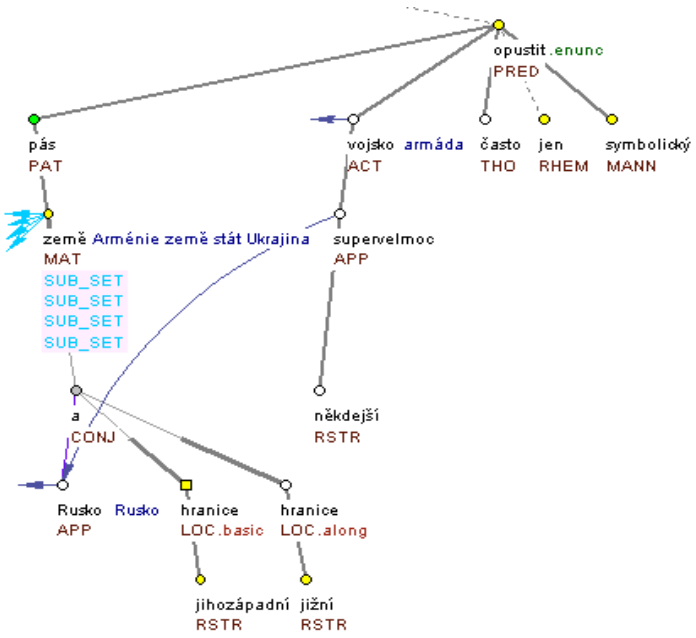

Fig. 1. Example of coreferential links (dark blue arrows) and bridging anaphora links (light blue arrows) depicted in TrEd.

The figure shows a tectogrammatical tree structure depicted in TrEd, representing the sentence: *Pás zemí podél jihozápadní a jižní hranice Ruska opustila vojska někdejší supervelmoci často jen symbolicky*. (*The army of the former superpower has often only symbolically left the line of countries along the southwest and south border of Russia*.)

5.1 Some helpful features of the annotation tool and a pre-annotation

In order to facilitate the task of the annotators, we implement some preliminary steps. The annotation can be made either on tectogrammatical trees, or directly in the text, preceding and following context being provided. All nodes that have the same lemma with the actual node, are underlined, so that the annotator can easily identify them in the text and decide if they should be marked as coreferential or not. Also, the already annotated relations are marked out by colors in the text. Single-word named entities and adjectives related to them (obtained from a list derived from a morphological synthesizer) are automatically pre-annotated, the annotators only need to check and correct the links. Finally, the automatic check of coreferential chains is implemented. If the annotator corefers the anaphor to the antecedent that is not the most recent one, the coreferential relation is automatically directed to the most recent one. Also, after a

deletion of a coreference, a possibly interrupted coreferential chain is automatically preserved (the annotator is asked to confirm it). This tool proves to be very useful especially for the cases where zero references participate in coreferential chains. It ensures that the annotator does not need to search for the last zero-pronoun node of the coreferential chain. E.g. in (9), the arrow leading from (9)c to (9)a is automatically re-directed to (9)b (provided that there already is a coreferential link from (9)b to (9)a).

Helena poprosila maminku$_a$, aby #PersPron[1]$_b$ na ni počkala. Maminka$_c$ však řekla, že #PersPron$_d$ nemůže.
(Helena asked [her] mother$_a$ to #PersPron$_b$ wait for her. However, the mother$_c$ said that [she] #PersPron couldn't.) (**9**)

5.2 The process of annotation

The annotation guidelines on the extended textual coreference and bridging relations were first drafted in 2007 [12]. After a series of annotation experiments, the annotation scheme is being applied large-scale, to the whole PDT corpus, by two instructed annotators, students of linguistics, whose portions of annotation are 1000 sentences/month. Table 1 shows the amount of data annotated so far.

Table 1. Annotation statistics

number of annotated documents	755
total number of sentences (in the annotated documents)	11,533
total number of words	193,386
total number of tectogrammatical nodes (excluding the technical roots)	156,928
number of newly annotated co-referring nodes (the textual coreference and the bridging anaphora)	20,141
number of originally annotated co-referring nodes (pronominal textual coreference)	5,191
number of all co-referring nodes (including the originally annotated pronominal coreference)	25,332
% of co-referring nodes	16 %
% of PDT 2.0 already annotated with the extended textual coreference and the bridging anaphora	24 %

In Figure 2, we present the proportion of different types of coreferential and bridging relations in the current annotation in PDT[2]. TK_0 is used for textual coreference of specific NPs, TK_NR for textual coreference of generic NPs, other abbreviations are believed to be self-explaining[3].

1 *#PersPron* here is the zero personal pronoun of the 3rd person

2 Including the originally annotated textual coreference in PDT 2.0.

3 In Fig.2, the bridging type bridging_ANAF is not present. The reason is that the type was newly defined and the statistics for it are not yet ready.

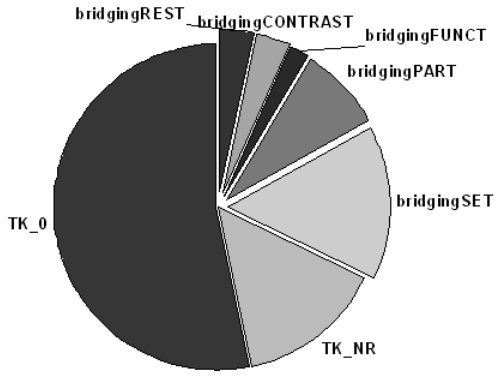

Fig. 2. Proportion of the types of relations in the annotations

5.3 Evaluation

The inter-annotator agreement has been measured four times (once every two months), on the total overlap of 280 sentences (40+40+100+100), and it was calculated separately for bridging anaphora and (the newly annotated) textual coreference using the F_1-measure on arguments of the relations and on arguments of the relations along with their types. Cohen's κ [1] was used to calculate agreement on the types of relations where the annotators agreed on the arguments. In Table 2, the results of the four performed measurements are presented.

Table 2. Evaluation of the inter-annotator agreement

	links textual coreference (F_1)	links textual coreference + types (F_1)	textual coreference types only (κ)	links bridging anaphora (F_1)	links bridging anaphora + types (F_1)	bridging anaphora types only (κ)
1st measurement (40 sent., 3 files)	0.76	0.67	0.54	0.49	0.42	0.79
2nd measurement (40 sent.,1 file)	0.64	0.41	0.33	0.52	0.52	1
3rdmeasurement (100 sent., 1 file)	0.80	0.68	0.67	0.59	0.57	0.88
4thmeasurement (100 sent., 2 files)	0.69	0.65	-0.02	0.42	0.39	0.93

Since we have only one stage of the annotation and use F_1-measure, it is not possible to directly compare the results with other similar works, which usually use two-level annotation scenario (selecting markables being the first step). Nevertheless, we can draw several observations from Table 2.

It is clear that choosing arguments of the bridging anaphora is a difficult task (the fourth column). In any case, it is more difficult than choosing arguments of the textual coreference (the first column). On the other hand, once the annotators agree on the arguments of the bridging anaphora, they very well agree on the type of the relation (κ on types is close to 0.8 or greater – the sixth column), while the agreement on type of the textual coreference is much worse (κ ranging usually from 0.33 to 0.67 – the third column); in the fourth measurement however, the agreement on the type of the textual coreference, measured by Cohen's κ, is even negative, which means that it is worse than random. It is caused by the fact that in the fourth measurement, there were about 100 coreferential links of the type 0 and only about 2 links of the type NR in the documents. In the previous measurements, the distribution of the types was much more even. Cohen's κ may not be a reliable measure in such a case. The F_1-measure on arrows and types (the second column) probably offers more informative numbers. It is also clear that the results very much depend on the text. See Section 6 for a detailed discussion of the inter-annotator (dis-)agreement.

6 Discussion of the inter-annotator agreement

Being not properly comparable to the existing well-known schemes of anaphoric annotation, the inter-annotation agreement is also greatly affected by parameters of the text as a whole. Short texts are generally far less demanding for their interpretation than longer texts of 20 to 120 sentences. Agreement is getting more difficult, the more complex the judgments required of the annotators are. So, the longer a chain is, the less likely that all annotators include all mentions in it. Also, the abstraction degree plays the crucial role for the results of the inter-annotator agreement (see Section 7).

A detailed study of the texts annotated by both annotators revealed several sources of typical errors. Typically, they are very similar to those presented in [18]. In the following paragraphs, some of them are discussed.

6.1 Different understanding of the content

This is probably the most frequent source of annotation disagreement. Annotators understand the anaphor as referring to different antecedents in the text, e.g. in (10):

Tak je i knížka koncipována. V každé kapitole se mluví o určitém problému. Je tam [coref_text with *knížka* or with *kapitola*] *v podstatě konkrétní návod.*
(That is the conception of the book. Every chapter discusses a particular problem. There are actually specific instructions there [coref_text with *the book* or with *chapter*]*.)* (**10**)

This kind of problem is not likely to be decided. The implementation of the ambiguity resolution strategies could possibly help, however also not perfectly. In many cases the annotator does not notice the existing ambiguity, annotating his subjective understanding of the text, the ambiguity being revealed only by the comparison of more annotation possibilities.

6.2 Choosing between bridging and generic textual coreference

With long (often hyperthematic) anaphoric chains of generic references, it is sometimes very difficult to decide between the textual coreference of a generic (NR) type and a the bridging relation (mostly of SUBSET type). The case in point is e.g. (11):

A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče$_a$*. V této knize je poučení, jak snášejí děti rozvod a jak na něj reagují, a návod, jak se mají rodiče*$_b$ [coref_text or bridging relation with *rodiče*] *chovat, aby se utrpení dětí snížilo*.
(After the book had been already written, it was clear, that it is quite useful for parents$_a$ *too. The book contains explanations how children go through divorce, how they react to it, and the instructions, how parents*$_b$ [coref_text or bridging relation with *parents*] *should behave to minimize the suffering of their children.)* **(11)**

In (11), the two NPs *rodiče* (*parents*) are used. In (11)a, the NP refers generically to the whole scope of parents; in (11)b, it refers to the divorced parents, but also generically. Two different decisions are made by the annotators – the annotator A related (11)b to (11)a as bridging SUBSET, the annotator B chose the generic textual coreference. Strictly judged, the second decision is more precise. However, this decision is very unlikely to be ever decided by any automatic method, nor seems to be of great importance for the text cohesion. On the other hand, the decision of the annotator B is quite counter-intuitive: the set of all parents is not coreferential with the set of the divorced parents.

In the PDT texts, such relations are very common, being usually also quite long, up to 30-element quasi-coreferential chains [12]. The problem in the inter-annotator agreement in this case is due to the annotation scheme and it could be solved by formulating more precise rules. However, this task proved to be very complicated. The case of generic NPs and NPs with other types of generic reference does not allow clear classification – every new context poses a new problem, not yet classified. The text cohesion in such contexts is not based on coreferential, but on other contextual relations (e.g. anaphora, associative relations etc.), so the language proposes no intuitive coreference annotation scheme in this case. Choosing the right convention, in order to achieve the higher inter-annotator agreement, is not easy. We cannot afford to follow both aims, the aim of automatic processing (information extraction, machine learning, etc.) and the aim of linguistic research.

6.3 Other causes of disagreement

Other frequent problems causing inter-annotator disagreement are e.g. the decision, whether the relation is to be annotated at all (more often for generic NPs), the selection of the antecedent for bridging relations, the decision between textual coreference to the root of the sentence or the special coreferential relation to the text segment (*coref_segm*), the decision for more intuitive coreferences, which contradict the postulated principles, etc. The disagreement is also increased by the way of measuring the inter-annotator agreement. In Figure 3, the annotator A creates a coreference between the two outer nodes, while the annotator B adds also the third node inbetween into the chain. Measuring the agreement between the annotators on the arguments of each arrow, we get a complete disagreement here, which we may or may not want to get.

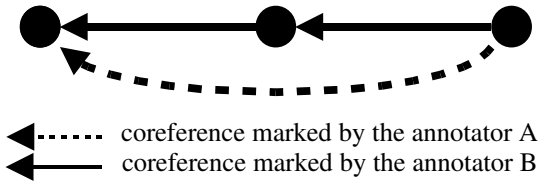

Fig. 3. An inter-annotator disagreement caused by a missing node in a coreferential chain

It is a well-known issue of measuring the inter-annotator agreement in the area of anaphora annotation, which can be solved using a more sophisticated approach, e.g. by applying Krippendorff's α [7] or another weighted coefficient on sets of nodes in the coreferential chains [15]. We plan to implement this approach later in the project, while at present we are more interested in detecting and studying detailed disagreements between the annotators, in order to detect problematic cases and improve the annotation instructions.

7 Conclusion and open questions

We have presented the ongoing project of the annotation of the extended textual coreference and the bridging anaphora in the Prague Dependency Treebank. We discussed in detail annotation guidelines and the typology of coreference relations and the bridging anaphora, and compared it with similar projects. Preliminary tests of the inter-annotator agreement were presented and we elaborated on various reasons for disagreement in the annotation. In the rest of this concluding section, let us present a few open questions.

The first phase of the coreference annotation process has revealed several problematic cases concerning annotation of anaphoric relations in Czech.

The most problematic aspect in annotating textual coreference concerns abstract nouns. Given that in some cases such NPs are clearly coreferential and anaphoric, we

cannot exclude them from the annotation. However, there exist much more cases, in which the decision for postulation of coreference is not so definite, sometimes appearing to be quite redundant, as e.g. in (12).

Míra nezaměstnanosti by se měla vyvíjet protikladně, než ve standardní ekonomice. […] růst nezaměstnanosti v letech 1991-1993 značně zaostal za poklesem HDP. Pokračující privatizace a restrukturalizace si však vynutí zvýšení míry nezaměstnanosti z 3,5% koncem roku 1993 na 5-6% ke konci příštího roku.
(The level of the unemployment should be developing oppositely to the standard economy. [...] The growth of the unemployment in 1991-1993 staid well behind the fall of GDP. However, the continuing privatization and restructuralisation will cause the increase in the level of unemployment from 3.5% at the end of the 1993 year to 5-6% at the end of the next year.) (**12**)

The following questions arise by deciding the annotation of cases like (12): should we annotate such cases at all? If we annotate them, what kind of coreference type is that (specific or generic coreference)? For the time being, we annotate relations between abstract nouns as generic coreference (*coref_text*, type 0), in order to be able to exclude them if needed. But there still remains the problem to distinguish between abstract and concrete nouns, the boundary being rather gradual (such nouns as *zisk* (*profit*), etc.).

There are some other open questions left, such as annotation of coreference in prepositional phrases, annotation of complex nouns, etc., which are solved using formal conventions.

Acknowledgements

This work was funded in part by the Companions project (www.companions-project.org), sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, as well as by the Czech Science Foundation (GAČR 405/09/0729), and by the Czech Ministry of Education (grant MSM-0021620838).

References

1. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46 (1960)
2. Hajič, J. et al.: Prague Dependency Treebank 2.0.CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia (2006)
3. Clark Herbert H.: Bridging, In Johnson-Laird, P. and P.C.Wason, editors, Thinking: Readings in Cognitive Science. Cambridge University Press, Cambridge (1977)
4. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R.: The Automatic Content Extraction (ACE) program - Tasks, data, and evaluation. In

Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004) (2004)
5. Hirschman, L.: MUC-7 coreference task definition. Version 3.0 (1997)
6. Krasavina O., Chiarcos, O.: PoCoS – Postdam Coreference Scheme. In Proc. of ACL-2007. Prague (2007)
7. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, chapter 12. Sage, Beverly Hills, CA (1980)
8. Kučová L., Hajičová E.: Coreferential Relations in the Prague Dependency Treebank. In 5th Discourse Anaphora and Anaphor Resolution Colloquium. Ediçôes Colibri (2004)
9. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z., Kučová, L.: Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka (t-layer annotation guidelines). Technical Report TR-2005-28, ÚFAL MFF UK, Prague (2005)
10. Mladová, L., Zikánová, Š., Hajičová, E.: From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In Proc. of LREC-2008 (2008)
11. Müller, C., Strube, M.: Multi-level annotation in MMAX. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, 198-207, Sapporo, Japan (2003)
12. Nedoluzhko, A.: Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závoslostním korpusu. (Annotating extended coreference and bridging relations in PDT. Technical report.). Prague (2007)
13. Orasan, C.: PALinkA: A highly customisable tool for discourse annotation. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue (2003)
14. Pajas, P., Štěpánek, J.: 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 (2008)
15. Passoneau, R. J.: Computing reliability for coreference annotation. In Proceedings of LREC, volume 4, pp. 1503-1506, Lisbon (2004)
16. Poesio, M.: The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In Proc. of SIGDIAL, Boston, April (2004)
17. Poesio, M., Artstein, R.: Anaphoric annotation in the ARRAU corpus. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008).
18. Recasens, M. (2008) Towards Coreference Resolution for Catalan and Spanish. Master Thesis. University of Barcelona (2008)
19. Recasens, M., Martí, A., Taulé, M.: Text as Scene: Discourse Deixis and Bridging Relations. Procesamiento del Lenguaje Natural, 39:205-212. Sevilla, Spain (2007)