A COREFERENTIALLY ANNOTATED CORPUS AND ANAPHORA RESOLUTION FOR CZECH

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz),
Mírovský J. (mirovsky@ufal.mff.cuni.cz),
Novák M. (mnovak@ufal.mff.cuni.cz)

Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

The paper presents an overview of a finished project focused on annotation of grammatical, pronominal and extended nominal coreference and bridging relations in the Prague Dependency Treebank (PDT 2.0). We give an overview of existing similar projects and their interests and compare them with our project. We describe the annotation scheme and the typology of coreferential and bridging relations and give the statistics of these types in the annotated corpus. Further we give the final results of the inter-annotator agreement with some explanations. We also briefly present the anaphora resolution experiments trained on the coreferentially annotated corpus and the future plans.

Keywords: anaphora, annotation, bridging relations, coreference, coreference resolution

1. Introduction

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. The Prague Dependency Treebank (PDT 2.0) (Jan Hajič et al., 2006) is a manually annotated corpus of Czech. The texts are annotated in 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 this paper we present an overview of the projects of annotating different types of coreference and bridging relations in the Prague Dependency Treebank, speak about the results of inter-annotator agreement and summarise some anaphora resolution experiments made on Czech data.

Section 2 describes the state of the art concerning annotating, analysing and using coreferentially annotated corpora. Section 3 gives a short overview of the types of coreference and bridging relations annotated in PDT. In Section 4, we give the statistics and discuss the results of inter-annotator agreement. Section 5 describes some anaphora resolution experiments that were made using the Czech coreferentially annotated data. We make conclusions in section 6.

2. PDT coreference and similar projects

The experiments on anaphora resolution, referential choice prediction, etc. are made using the annotated corpora for coreference. There are a number of different large-scale annotated corpora for coreference and anaphoric relations on which the experiments for coreference resolution are made. The largest annotated corpora for English include MUC (Hirschman and Chinchor, 1997), ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007), GNOME (Poesio, 2004), ARRAU (Poesio and Artstein, 2008). The coreference annotations for other languages than English are more limited. The most well-known corpora including anaphoric informations are AnCora (Recasens and Marti, 2009) for Spanish and Catalan, VENEX (Poesio et al., 2004) for spoken and written Italian, the Italian Live Memories Corpus (Rodríguez et al., 2010), TüBA-DZ Treebank (Hinrichs et al., 2004) and Postdam Commantary Corpus (Stede, 2008; Krasavina and Chiarcos, 2007) for German, PdiTB (Poláková et al., 2012) etc.

Determining coreference is a highly complicated task, and even between human annotators there is a lot of disagreement leading to a relatively low number of inter-annotator agreement, especially concerning nominal coreference and bridging relations. The cases of vagueness and referential ambiguity were a subject of a rich discussion in computational linguistics and anaphoric community during the last few years. There were discussed such topics as e.g. justified sloppiness hypothesis in Poesio et al. 2006, the reliability of anaphoric annotation in Poesio and Artstein 2005, examples and reasons for vagueness and referential ambiguity in Versley 2008, so-called near-identity relation in Recasens et al. 2010. Some discussion on ambiguous cases of coreference and the reasons for inter-annotator disagreement for Czech were presented in Nedoluzhko 2010.

3. Types of coreference and bridging relations annotated in PDT

In PDT 2.0, two types of coreference (grammatical and textual) and six types of bridging relations have been annotated. The **grammatical coreference** typically occurs within a single sentence, the antecedent being able to be derived on the basis of grammar rules of a given language. It includes relative pronouns, verbs of control, reflexive pronouns, reciprocity and verbal complements. The detailed description of the types of grammatical coreference and the examples may be found in Mikulová et al. 2006.[1] **Textual coreference** is generally taken to mean the use of various linguistic means (pronouns, synonyms, generalizing nouns etc.) which function as anaphoric (occasionally cataphoric) reference devices. This reference is not expressed by grammatical means alone, but also via context. As for textual coreference in PDT, it has been annotated in two time periods. At first, the so-called pronominal textual coreference was manually annotated. It was restricted to cases in which a demonstrative

[1] The resumed typology of grammatical coreference in PDT was also presented at DIALOG in Nedoluzhko 2009.

this or an anaphoric pronoun of the 3rd person, also in its zero form, are used (Kučová and Hajičová, 2004). Afterwards, the annotation of textual coreference was extended to cases where the anaphor is expressed by other means such as full noun phrases (definite descriptions, repetitions, synonyms etc.), adverbs (there, then etc.) and some types of numerals and pronouns neglected in the first stage. This stage of the project was called the Extended Textual Coreference and described in detail during the annotation period in (Nedoluzhko et al. 2009; Nedoluzhko, 2011; Nedoluzhko and Mírovský, 2011). Annotation of extended textual coreference and bridging relation is a project related to PDT 2.5 (Bejček et al., 2011), which is a revised, updated and extended version of PDT 2.0.

The textual coreference is further classified into two types — coreference of NPs with specific (type SPEC) or generic (type GEN) reference. Compare examples (1) and (2):

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

(2) *Lions live in a forest. They are not vegetarians* [type GEN].

Special cases of textual coreference. Two special cases of reference are 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 where the antecedent is understood by inferencing from a broader co-text. This kind of relation has no explicitly marked antecedent, it just proves 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.

For the **bridging relations**, the following types are distinguished: part-of relation (*room — ceiling*), set — subset (*students — some students*) and FUNCT (*trainer — football team*) traditional relations, CONTRAST for coherence relevant discourse opposites, ANAF for explicitly anaphoric relations without coreference and the further underspecified group REST. The more detailed description of types can be found for example in Nedoluzhko and Mírovský 2011.

4. Statistics and inter-annotator agreement

By the end of 2011, the whole PDT data was annotated for coreference and bridging relations (see Nedoluzhko et al. 2011).[2] Table 1 shows the statistics of the annotated data.

[2] The completed and corrected version was published together with the annotation of discourse relations in the Prague Discourse Treebank in 2012 (see Poláková et al. 2012).

Table 1. Statistics of the annotated data

Total number of sentences (in the annotated documents)	49,431
Total number of tokens	833,195
Number of coreferring nodes — grammatical coreference	23,272
Number of coreferring nodes — textual coreference	86,349
Number bridging relations	32,171
% of co-referring nodes	17,6%

As for the distribution of types of textual coreference and bridging relations, the proportion is represented in Table 2:

Table 2. The distribution of types of textual coreference and bridging relations

Type	**Number**
Textual coreference (specific)	20,243 (pronouns) + 50,593 (nouns) = 70,836
Textual coreference (generic)	3,095 (pronons) + 12,418 (nouns) = 15,513
All textual coreference links	86,349
All bridging links	32,171

As seen from the table, textual coreference makes the significant majority of the annotated relations and inside the group of textual coreference the coreference of specific noun phrases significantly prevail. The reason for relatively low percentage of bridging relations may be mainly the small number of types and their precise delimitating (even for annotation of the bridging relation of type REST, very precise rules were set). As for the significant dominance of textual coreference between specific noun phrases over generic ones, the reasons are mainly empiric. Also postulating coreference between generic noun phrases is a much more complicated task than coreferring specific noun phrases, so in most existing projects it is excluded from the annotation of coreference (Poesio, 2004; Recasens, 2010 etc.).

We have measured the inter-annotator agreement in the annotation of coreference and bridging anaphora in PDT on a small part of the data that had been annotated in parallel by two annotators. To evaluate the agreement, we have used the chain-based F1-measure, a simple ratio, and Cohen's κ (Cohen, 1960). The chain-based F1-measure has been used for measuring the agreement on the recognition of a coreference or bridging relation, a simple ratio and Cohen's κ have been used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation.

In the chain-based measure, we consider the annotators to be in agreement on recognizing a coreference or bridging relation if the two nodes connected by an arrow by one of the annotators have also been connected by the other annotator; coreference chains are taken into account, i.e. it is sufficient for the agreement if the arrow starts in or goes to a node that is coreferentially connected (possibly transitively) with the node used for the relation by the other annotator.

Table 3. Results of the inter-annotator agreement

Measurement	F1	Agreement on types	Kappa on types
All parallel data — coreference	0.72	0.90	0.73
All parallel data — bridging anaphora	0.46	0.92	0.89

Table 3 shows that the results for inter-annotator agreement are not particularly high. In our measurements and analyses of inter-annotator agreement, we observe the three main types of disagreement: (a) disagreement concerning the decision if the relation in question should be annotated as a coreference (or bridging) relation, (b) disagreement on choosing the antecedent and (c) disagreement in the type of the annotated relation. The reasons for relatively low numbers of inter-annotator discrepancies and the typology of disagreements with the examples were discussed in Nedoluzhko 2010.

5. Automatic experience on the annotated data

The main objective of our annotation effort has been to provide data for developing automatic techniques for resolution of anaphoric relations. PDT has served as a source of gold standard data for testing as well as a source of training data for tools utilizing machine learning methods.

Antecedents in grammatical coreference can be usually derived with high accuracy from grammatical rules. Nguy 2006 presented a set of rules for various types of grammatical coreference, achieving more than 90% F1-measure for every type.

In Nguy and Žabokrtský 2007, a rule-based system was employed to resolution of pronominal textual coreference. Higher complexity of this task affects the success rate which is substantially lower (74% F1-measure) than what can be reached in the task of grammatical coreference resolution. Applying machine learning methods, particularly perceptron ranking in Nguy et al. 2009, on the same task outperformed the rule-based method with F1-measure over 79%.

However, the features used in these experiments were extracted from the manually annotated tectogrammatical layer of PDT 2.0. Thus the system could take advantage of perfectly correct information on various linguistic attributes which are not available in a real-world scenario. In Bojar et al. 2012, the authors used the same perceptron ranker and the same feature set for training and testing, this time extracted from the automatically analyzed data though. Unreliability of information on tectogrammatical gender and number as well as uncertainty of presence of a subject omitted on the surface[3] resulted in a substantial drop in performance to 50% F1-measure.

It confirms that correct identification of an unexpressed subject and determination, whether it is anaphoric, is central to resolution of the zero variant of pronominal coreference. This and a corresponding issue in English — determination of whether a personal pronoun "it" is anaphoric — were addressed in the work of Veselovská et al. 2012

[3] Czech is a pro-drop language.

by a set of rules tested on Prague Czech-English Dependency Treebank 2.0 (PCEDT). Some of these rules made use of parallel nature of the treebank by providing information from the English side to facilitate identification of Czech unexpressed subjects.

Annotation work on the Extended Textual Coreference project encouraged research on noun phrase (NP) textual coreference resolution. Novák 2010 carried out the first experiments on NP coreference in Czech. The approach of maximum entropy ranking was further elaborated in Novák and Žabokrtský 2011, where authors compared systems based on classification and ranking approaches in machine learning. As a result, the best system achieves 44.4% F1-measure on coreference with specific reference. Novák 2010 also paid his attention on coreference with generic reference as well as bridging relations of the type PART. Despite the unsatisfying results, his work introduces a novel feature inspired by Hearst patterns (Hearst, 1992) that is supposed to capture a PART-WHOLE relation by exploiting a large morphologically annotated corpus.

Knowledge of anaphoric relations in a text can be crucial to solving more complex tasks. Multiple tools mentioned above have been integrated with a modular NLP framework Treex (Popel and Žabokrtský, 2010) that is used in various scenarios. For instance, the rules for resolving grammatical and pronominal textual coreference contribute on English to Czech translation in TectoMT system (Žabokrtský et al., 2008). In addition, some of these tools and their counterparts for English helped to form both sides of the automatically annotated Czech-English parallel corpus CzEng 1.0 (Bojar et al., 2011), consisting of over 15 million sentence pairs.

The overview of performance of the tools mentioned above can be found in Table 4.

Table 4. The overview of performance of the tools

Type of the task	Published	Data	Success rate
Grammatical coreference, verbs of control	Nguy 2006	PDT 2.0	91.5%
Grammatical coreference, reflexive pronouns	Nguy 2006	PDT 2.0	97.1%
Grammatical coreference, relative pronouns	Nguy 2006	PDT 2.0	99.6%
Grammatical coreference, reciprocity	Nguy 2006	PDT 2.0	94.7%
Pronominal coreference, rule-based	Nguy and Žabokrtský 2007	PDT 2.0	74.2%
Pronominal coreference, perceptron ranking, gold features	Nguy et al. 2009	PDT 2.0	79.4%
Pronominal coreference, perceptron ranking, system features	Nguy et al. 2009	PDT 2.0	50.3%
Identification of an anaphoric unexpressed subject, rule-based	Veselovská et al. 2012	PCEDT 2.0	61.5%
Identification of an anaphoric unexpressed subject, rule-based, expoliting English side	Veselovská et al. 2012	PCEDT 2.0	69.5%
NP coreference, maximum entropy ranking	Novák 2010	PDT 2.5	39.4%
NP coreference, perceptron ranking, improved features	Novák and Žabokrtský 2011	PDT 2.5	44.4%

6. Conclusion and future work

We presented the finished project of the Czech annotation of different types of coreference and bridging relations. We compared our project to other similar projects, gave the statistics of coreference and bridging types and the results for inter-annotator agreement. We briefly presented the anaphora resolution experiments trained on coreferentially annotated corpus.

At present, we are completing the annotation for the first and second person coreference. In future, other corpora for Czech (e.g. the Prague Dependency Treebank of Spoken Czech, Prague Czech-English Dependency Treebank) are to be supplied with some types of coreferential relations.

Acknowledgements

We gratefully acknowledge the support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875) and GAUK 4226/2011.

References

1. *Bojar, Ondřej; Žabokrtský, Zdeněk; Dušek, Ondřej; Galuščáková, Petra; Majliš, Martin; Mareček, David; Maršík, Jiří; Novák, Michal; Popel, Martin; Tamchyna, Aleš:* CzEng 1.0. Data, Charles University in Prague, UFAL, 2011.
2. *Bojar, Ondřej; Žabokrtský, Zdeněk; Dušek, Ondřej; Galuščáková, Petra; Majliš, Martin; Mareček, David; Maršík, Jiří; Novák, Michal; Popel, Martin; Tamchyna, Aleš:* The Joy of Parallelism with CzEng 1.0. In Proceedings of LREC 2012, İstanbul, 2012.
3. *Cohen, Jacob:* A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), 1960.
4. *Doddington, George; Mitchell, Alexis; Przybocki, Mark; Ramshaw, Lance; Strassel, Stephanie; Weischedel, Ralph:* The Automatic Content Extraction (ACE) program — tasks, data, and evaluation. In Proceedings of LREC 2004, Lisbon, 2004.
5. *Hajič, Jan et al.:* Prague Dependency Treebank 2.0.CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.
6. *Hearst, Marti A.:* Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th Conference on Computational linguistics — Volume 2, Nantes, France, 1992.
7. *Hinrichs, Erhard; Kübler, Sandra; Naumann, Karin; Telljohann, Heike; Trushkina, Julia:* Recent developments in linguistic annotations of the TüBa-D/Z treebank. In Proceedings of the Third Workshop on Treebanks and Linguistic Theories, Tübingen, 2004.
8. *Hirschman, Lynette; Chinchor, Nancy:* MUC-7 Coreference Task Definition — Version 3.0, 1997.
9. *Krasavina, Olga; Chiarcos, Christian:* PoCoS — Potsdam Coreference Scheme. In Proceedings of the Linguistic Annotation Workshop, Prague, 2007.

10. *Kučová, Lucie; Hajičová, Eva:* Coreferential Relations in the Prague Dependency Treebank. In Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium, S. Miguel, 2004.
11. *Mikulova Marie et al.:* Anotace na tektogramatické rovině Pražského závislostního korpusu. Referenční příručka. Technical report no. 2006/31, Charles University in Prague, UFAL, 2006.
12. *Nguy, Giang Linh:* Proposal of a set of rules for anaphora resolution in Czech. Master thesis, Charles University in Prague, 2006.
13. *Nguy, Giang Linh; Novák, Václav; Žabokrtský, Zdeněk:* Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In Proceedings of the SIGDIAL 2009 Conference, London, 2009.
14. *Nguy, Giang Linh; Žabokrtský, Zdeněk:* Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data. In Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium, Lagos, 2007.
15. *Nedoluzhko, Anna:* Coreferential relationships in text — comparative analysis of annotated data. In Papers from the Annual International Conference "Dialogue 2010" Issue 9 (16), Moscow, 2010.
16. *Nedoluzhko, Anna:* Razmetka koreferencii na sintaksičeski annotorovannom korpuse češskich tekstov. In Papers from the Annual International Conference "Dialogue 2009" Issue 8 (15), Moscow, 2009.
17. *Nedoluzhko, Anna; Mírovský, Jiří; Ocelák, Radek; Pergler, Jiří:* Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. In Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium, Goa, 2009.
18. *Nedoluzhko, Anna; Mírovský, Jiří; Pajas, Petr:* Annotation Tool for Extended Textual Coreference and Bridging Anaphora. In Proceedings of LREC 2010, Malta, 2010.
19. *Nedoluzhko, Anna; Mírovský, Jiří:* Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank. Technical report no. 2011/44, Charles University in Prague, UFAL, 2011.
20. *Nedoluzhko, Anna; Mírovský, Jiří, Hajičová, Eva, Pergler, Jiří, Ocelák, Radek:* Extended Textual Coreference and Bridging Relations in PDT 2.0. Data. Charles University in Prague, ÚFAL, 2011.
21. *Nedoluzhko, Anna:* Rozšířená textová koreference a asociační anaphora (Koncepce anotace českých dat v Pražském závislostním korpusu). ÚFAL, Praha, 2011.
22. *Novák, Michal:* Machine Learning Approach to Anaphora Resolution. Master thesis, Charles University in Prague, 2010.
23. *Novák, Michal; Žabokrtský, Zdeněk:* Resolving Noun Phrase Coreference in Czech. In Lecture Notes in Computer Science 7099, Springer-Verlag Heidelberg, 2011.
24. *Poesio, Massimo; Delmonte, Rodolfo; Bristot, Antonella; Chiran, Luminita; Tonelli, Sara:* The Venex corpus of anaphora and deixis in spoken and written Italian. Manuscript, 2004.

25. *Poesio, Massimo; Artstein, Ron:* The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, Ann Arbor, 2005.
26. *Poesio, Massimo; Sturt, Patrick; Artstein, Ron; Filik, Ruth:* Underspecification and Anaphora: Theoretical Issues and Preliminary Evidence. In Discource Processes 42(2), 2006.
27. *Poesio, Massimo:* The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In Proceedings of The 5th SIGdial Workshop on Discourse and Dialogue, Boston, 2004.
28. *Poesio, Massimo; Artstein, Ron:* Anaphoric annotation in the ARRAU corpus. In Proceedings of LREC 2008, Marrakech, 2008.
29. *Poláková, Lucie; Jínová, Pavlína; Zikánová, Šárka; Hajičová, Eva; Mírovský, Jiří; Nedoluzhko, Anna; Rysová, Magdaléna; Pavlíková, Veronika; Zdeňková, Jana; Pergler, Jiří; Ocelák, Radek:* Prague Discourse Treebank 1.0. Data, Charles University in Prague, ÚFAL, 2012.
30. *Popel, Martin; Žabokrtský, Zdeněk:* TectoMT: Modular NLP Framework. In Lecture Notes in Computer Science, Vol. 6233, Springer-Verlag Heidelberg, 2010.
31. *Pradhan, Sameer S.; Hovy, Eduard; Marcus, Mitch; Palmer, Martha; Ramshaw, Lance; Weischedel, Ralph:* Ontonotes: A unified relational semantic representation. In Proceedings of the International Conference on Semantic Computing, Washington DC., 2007.
32. *Recasens, Marta; Hovy, Eduard; Martí, M. Antònia:* A typology of near-identity relations for coreference (NIDENT). In Proceedings of LREC 2010, Valletta, 2010.
33. *Recasens, Marta; Martí, M. Antònia:* AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation, 2009.
34. *Rodríguez, Kepaj; Delogu, Francesca; Versley, Yannick; Stemle, Egon W.; Poesio, Massimo:* Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In Proceedings of LREC 2010, Valletta, 2010.
35. *Stede, Manfred:* Disambiguating Rhetorical Structure. In Research on Language and Computation, Vol. 6, Issue 3–4, Springer Netherlands, 2008.
36. *Versley, Yannick:* Vagueness and referential ambiguity in a large-scale annotated corpus. In Research on Language and Computation 6 (3–4), Springer Netherlands, 2008.
37. *Veselovská, Kateřina; Nguy, Giang Linh; Novák, Michal:* Using Czech-English Parallel Corpora in Automatic Identification of 'It'. In The Fifth Workshop on Building and Using Comparable Corpora, İstanbul, 2012.
38. *Žabokrtský, Zdeněk; Ptáček, Jan; Pajas, Petr:* TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, 2008.