Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution

Lucie Kučová and Zdeněk Žabokrtský *

Institute of Formal and Applied Linguistics, Charles University (MFF),
Malostranské nám. 25, CZ-11800 Prague, Czech Republic
{kucova,zabokrtsky}@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz

Abstract. The aim of this paper is two-fold. First, we want to present a part of the annotation scheme of the Prague Dependency Treebank 2.0 related to the annotation of coreference on the tectogrammatical layer of sentence representation (more than 45,000 textual and grammatical coreference links in almost 50,000 manually annotated Czech sentences). Second, we report a new pronoun resolution system developed and tested using the treebank data, the success rate of which is 60.4 %.

1 Introduction

Coreference (or co-reference) is usually understood as a symmetric and transitive relation between two expressions in the discourse which refer to the same entity. It is a means for maintaining language economy and discourse cohesion ([1]). Since the expressions are linearly ordered in the time of the discourse, the first expression is often called *antecedent*. Then the second expression (*anaphor*) is seen as 'referring back' to the antecedent. Such a relation is often called *anaphora*.[1] The process of determining the antecedent of an anaphor is called *anaphora resolution* (AR).

Needless to say that AR is a well-motivated NLP task, playing an important role e.g. in machine translation. However, although the problem of AR has attracted the attention of many researches all over the world since 1970s and many approaches have been developed (see [2]), there are only a few works dealing with this subject for Czech, especially in the field of large (corpus) data.

The present paper summarizes the results of studying the phenomenon of coreference in Czech within the context of the Prague Dependency Treebank 2.0 (PDT 2.0).[2] PDT 2.0 is a collection of linguistically annotated data and documentation and is based on the theoretical framework of Functional Generative Description (FGD). The annotation scheme of the PDT 2.0 consists of

* The research reported on in this paper has been supported by the grant of the Charles University in Prague 207-10/203329 and by the project 1ET101120503.

[1] Unfortunately, these terms tend to be used inconsistently in literature.

[2] PDT 2.0 is to be released soon by the Linguistic Data Consortium.

three layers: morphological, analytical and tectogrammatical. Within this system, coreference is captured at the tectogrammatical layer of annotation.

2 Theoretical Background

In FGD, the distinction between grammatical and textual coreference is drawn ([6]). One of the differences is that (individual subtypes of) grammatical coreference can occur only if certain local configurational requirements are fulfilled in the dependency tree (such as: if there is a relative pronoun node in a relative clause and the verbal head of the clause is governed by a nominal node, then the pronoun node and nominal node are coreferential), whereas textual coreference between two nodes does not imply any syntactic relation between the nodes in question or any other constraint on the shape of the dependency tree. Thus textual coreference easily crosses sentence boundaries.

Grammatical Coreference. In the PDT 2.0, grammatical coreference is annotated in the following situations (see a sample tree in Fig. 1):[3] (i) relative pronouns in relative clauses, (ii) reflexive and reciprocity pronouns (usually coreferential with the subject of the clause), (iii) control (in the sense of [7]) – both for verbs and nouns of control.

Textual Coreference. For the time being, we concentrate on the case of textual coreference in which a demonstrative or an anaphoric pronoun (also in its zero form) are used.[4] The following types of textual coreference links are special (see a sample tree in Fig. 2):[5]

- a link to a particular node if this node represents an antecedent of the anaphor or a link to the governing node of a subtree if the antecedent is represented by this node plus (some of) its dependents:[6] *Myslíte, že rozhodnutí NATO, zda se [**ono**] rozšíří, či nikoli, bude záviset na postoji Ruska?* (Do you think that the decision of NATO whether [**it**] will be enlarged or not will depend on the attitude of Russia?)
- a specifically marked link (segm) denoting that the referent is a whole segment of text, including also cases, when the antecedent is understood by inferencing from a broader co-text: *Potentáti v bance koupí za 10, prodají si za 15.(. . .) Odhaduji, že do 2 let budou schopni splatit bance dluh a třetím*

[3] We only list the types of coreference in this paper; detailed linguistic description will be available in the documentation of the PDT 2.0.

[4] With the demonstrative pronoun, we consider only its use as a noun, not as an adjective; we do not include pronouns of the first and second persons.

[5] Besides the listed coreference types, there is one more situation where coreference occurs but is difficult to be identified and no mark is stored into the attributes for coreference representation. It is the case of nodes with tectogrammatical lemma #Unsp (unspecified); see [9]. Example: *Zmizení tohoto 700 kg těžkého přístroje hygienikům ohlásili (**Unsp**) 30. června letošního roku.* (Lit.: The disappearance of the medical instrument weighing 700 kg to hygienists[**they**] announced on June 30th this year.)

[6] This is also the way how a link to a clause or a sentence is being captured.

rokem už budou dělat na sebe. A na práci najmou jen schopné lidi. Kdo ***to*** *pochopí, má náskok.* (The big shots buy in a bank for 10 and sell for 15. (...) I guess that within two years they will be able to pay back the debt to the bank and in the third year they will work for themselves. And they will hire only capable people, it will be in their best interest. Those who understand **this**, will have an advantage.)

– a specifically marked link (exoph) denoting that the referent is "out" of the co-text, it is known only from the situation: *Následuje dramatická pauza a pak již vchází* ***On*** *nebo* ***Ona****.* (Lit. (there) follows dramatic pause and then already enters **He** or **She**.)

3 Annotated Data

Data Representation. When designing the data representation on coreference links, we took into account the fact that each tectogrammatical node is equipped with an identifier which is unique in the whole PDT. Thus the coreference link can be easily captured by storing the identifier of the antecedent node (or a sequence of identifiers, if there are more antecedents for the same anaphor) into a distinguished attribute of the anaphor node. We find this 'pointer' solution more transparent (and – from the programmer's point of view – much easier to cope with) than the solutions proposed in [3] or [4].

At present, there are three node attributes used for representing coreference: (i) coref_gram.rf – identifier or a list of identifiers of the antecedent(s) related via grammatical coreference; (ii) coref_text.rf – identifier or a list of identifiers of the antecedent(s) related via textual coreference; (iii) coref_special – values segm (segment) and exoph (exophora) standing for special types of textual coreference.

We used the tree editor TrEd developed by Petr Pajas as the main annotation interface.[7] More details concerning the annotation environment can be found in [8]. In this editor (as well as in Figures 1 and 2 in this paper), a coreference link is visualized as a non-tree arc pointing from the anaphor to its antecedent.

Quantitative Properties. PDT 2.0 contains 3,168 newspaper texts annotated at the tectogrammatical level. Altogether, they consist of 49,442 sentences with 833,357 tokens (summing word forms and punctuation marks). Coreference has been annotated manually (disjunctively[8]) in all this data. After finishing the manual annotation and post-annotation checks and corrections, there are 23,266 links of grammatical coreference (dominating relative pronouns as the anaphor – 32 %) and 22,365[9] links of textual coreference (dominating personal and possessive pronouns as the anaphor – 83 %), plus 505 occurrences of segm and 120 occurrences of exoph).

[7] `http://ufal.mff.cuni.cz/~pajas`

[8] Independent parallel annotation of the same sentences were performed only in the starting phase of the annotation, only as long as the annotation scheme stabilized and reasonable inter-annotator agreement was reached (see [8])

[9] Similarity of the numbers of textual and grammatical coreference links is only a more or less random coincidence. If we would have annotated also e.g. bridging anaphora, the numbers would be much more different.

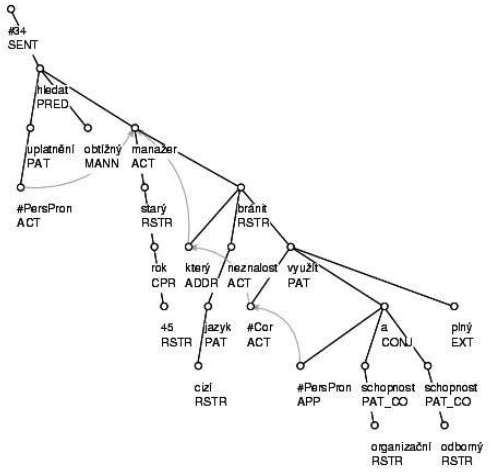

Fig. 1. Simplified PDT sample with various subtypes of grammatical coreference. The structure is simplified, only tectogrammatical lemmas, functors, and coreference links are depicted. The original sentence is '*Obtížněji hledají své uplatnění manažeři starší 45 let, kterým neznalost cizích jazyků brání plně využít své organizační a odborné schopnosti.*' (Lit.: More difficultly search their self-fulfillment manages older than 45 years, to which unknowledge of foreign languages hamper to use their organization and specialized abilities).

4 Experiments and Evaluation of Automatic Anaphora Resolution

In [8] it was shown that it is easy to get close to 90 % precision when considering only grammatical coreference.[10] Obviously, textual coreference is more difficult to resolve (there are almost no reliable clues as in the case of grammatical coreference). So far, we attempted to resolve only the textual coreference links 'starting' in nodes with tectogrammatical lemma #PersPron. This lemma stands for personal (and personal possessive) pronouns, be they expressed on the surface (i.e., present in the original sentence) or restored during the annotation of the tectogrammatical tree structure.

We use the following procedure (numbers in parentheses were measured on the training part of the PDT 2.0):[11] For each detected anaphor (lemma #PersPron):

[10] This is not surprising, since in the case of grammatical coreference most of the information can be derived from the topology and basic attributes of the tree (supposing that we have access also to the annotation of morphological and analytical level of the sentence). However, it opens the question of redundancy (at least for certain types of grammatical coreference).

[11] The procedure is based mostly on our experience with the data. However, it undoubtedly bears many similarities with other approaches ([2]).

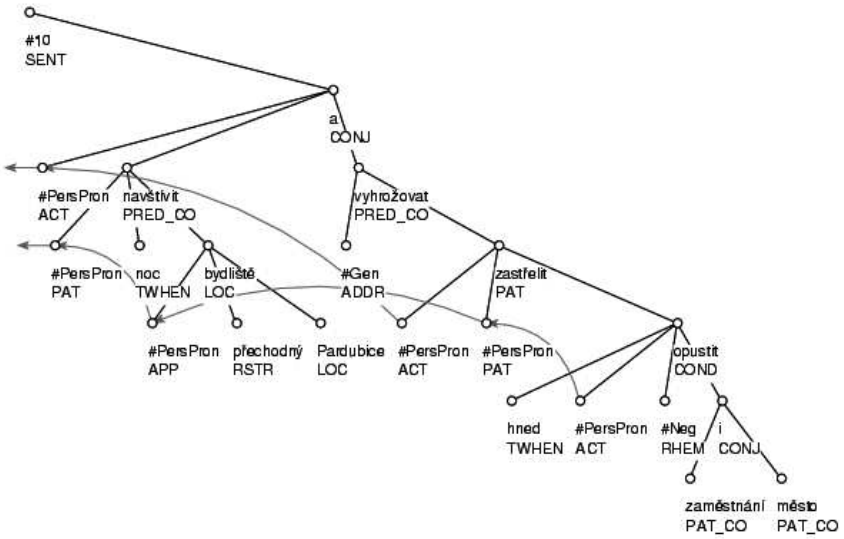

Fig. 2. Simplified PDT sample containing two textual coreference chains. The original sentence is '*Navštívil ji v noci v jejím přechodném bydlišti v Pardubicích a vyhrožoval, že ji zastřelí, pokud hned neopustí zaměstnání i město.*' (Lit.: [He] visited her in night in her temporary dwelling in Pardubice and threatened [her] that [he] will shoot her if [she] instantly does not leave her job and city.).

– First, an initial set of antecedent candidates is created: we used all nodes from the previous sentence and current sentence (roughly 3.2 % of correct answers disappear from the set of candidates in this step).
– Second, the set of candidates is gradually reduced using various filters: (1) candidates from the current sentence not preceding the anaphor are removed (next 6.2 % lost), (2) candidates which are not semantic nouns (nouns, pronouns and numeral with nominal nature, possessive pronouns, etc.), or at least conjunctions coordinating two or more semantic nouns, are removed (5.6 % lost), (3) candidates in subject position which are in the same clause as the anaphor are removed, since the anaphor would be probably expressed by a reflexive pronoun (0.7 % lost) (4) all candidates disagreeing with the anaphor in gender or number are removed (3.7 % lost), (5) candidates which are parent or grandparent of the anaphor (in the tree structure) are removed (0.6 % lost), (6) if both the node and its parent are in the set of candidates, then the child node is removed (1.6 % lost), (7) if there is a candidate with the same functor with anaphor, then all candidates having different functor are removed (3.4 % lost), (8) if there is a candidate in a subject position, then all candidates in different than subject positions are removed (2.4 % lost),
– Third, the candidate is chosen from the remaining set which is (linearly) the closest to the given anaphor (12.5 % lost).

When measuring the performance only on the evaluation-purpose part of the PDT 2.0 data (roughly 10 % of the whole), the final success rate (number of correctly resolved antecedents divided by the number of pronoun anaphors) is 60.4 %.[12]

The whole system consists of roughly 200 lines of Perl code and was implemented using ntred[13] environment for accessing the PDT data. The question of speed is almost irrelevant: since the system is quite straightforward and fully deterministic, ntred running on ten networked computers needs less than one minute to resolve all #PersPron node in PDT.

5 Final Remarks

We understand coreference as an integral part of a dependency-based annotation of underlying sentence structure which prepares solid grounds for further linguistic investigations. It proved to be useful in the implemented AR system, which profits from the existence of the tectogrammatical dependency tree (and also from the annotations on the two lower levels).

As for the results achieved by our AR system, to our knowledge there is no other system for Czech reaching comparable performance and verified on comparably large data.

References

1. Halliday M. A. K., Hasan, R.: Cohesion in English, Longman, London (1976)
2. Mitkov, R.: Anaphora resolution. Longman, London (2001)
3. Plátek, M., Sgall, J., Sgall, P.: A Dependency Base for a Linguistic Description. In: Sgall, P.(ed.): Contributions to Functional Syntax, Semantics and Language Comprehension. Academia, Prague (1984) 63-97
4. Hajičová, E., Panevová J., Sgall, P.: Coreference in Annotating a Large Corpus. In: Proceedings of LREC 2000, Vol. 1. Athens, Greece (2000) 497-500
5. Hajičová, E., Panevová, J., Sgall, P. Manuál pro tektogramatické značkování. Technical Report ÚFAL-TR-7 (1999)
6. Panevová, J.: Koreference gramatická nebo textová? In: Banys, W., Bednarczuk, L., Bogacki, K. (eds.): Etudes de linguistique romane et slave, Krakow (1991)
7. Panevová, J.: More Remarks on Control. In: Prague Linguistic Circle Papers, Benjamin Publ. House, Amsterdam – Philadelphia (1996) 101-120
8. Kučová L., Kolářová V., Pajas P., Žabokrtský Z., and Čulo O.: Anotování koreference v Pražském závislostním korpusu. Technical Report of the Center for Computational Linguistics, Charles University, Prague (2003)
9. Kučová L., Hajičová E. 2004), Coreferential Relations in the Prague Dependency Treebank. Presented at 5th Discourse Anaphora and Anaphor Resolution Colloquium, San Miguel, Azores (2004)
10. Barbu C., Mitkov, R.: Evaluation tool for rule-based anaphora resolution methods. In: Proceedings of ACL'01, Toulouse, France (2001) 34-41

[12] For instance, the results in pronoun resolution in English reported in [10] was also around 60 %.

[13] `http://ufal.mff.cuni.cz/~pajas`