Eva Hajičová, Jarmila Panevová and Petr Sgall

COREFERENCE IN ANNOTATING A LARGE CORPUS

Faculty of Mathematics and Physics, Charles University

Malostranské nám. 25, 11800 Prague 1, Czechia

hajicova,panevova,sgall}@ufal.mff.cuni.cz

ABSTRACT

The Prague Dependency Treebank (PDT) is a part of the Czech National Corpus, annotated with disambiguated structural descriptions representing the meaning of every sentence in its environment. To achieve that aim, it is necessary i.a. to make explicit (at least some basic) coreferential relations within the sentence boundaries and also beyond them. The PDT scenario includes both automatic and 'manual' procedures; among the former type, there is one that concerns coreference, indicating the lemma of the subject in a specific attribute of the label belonging to a node for a reflexive pronoun, and assigning the deleted nodes in coordinated constructions the lemmas of their counterparts in the given construction. 'Manual' operations restore nodes for the deleted items mostly as pronouns.

The distinction between grammatical and textual coreference is reflected. In order to get a possibility of handling textual coreference, specific attributes reflect the linking of sentences to each other and to the context of situation, and the development of the degrees of activation of the 'stock of shared knowledge' will be registered in so far as they are derivable from the use of nouns in subsequent utterances in a discourse.

1. OVERVIEW OF THE ANNOTATION PROCEDURE

1.1. The units of annotation in the **Prague Dependency Treebank** (PDT) are sentences as occurring in the texts in the Czech National Corpus, and the human annotators are instructed to assign every sentence a (disambiguated) structural description according to the meaning of the

sentence in its environment. In the manual phase, the annotators are helped by a 'user-friendly' software that makes it possible to work with diagrammatic shapes of the trees.

Several parts of the tagging procedure can be formulated as general steps, carried out automatically (see Hajič 1998, Hajičová 1998). One of these parts follows after the dependency structure of the sentence (the nodes of the dependency tree and the syntactic relations indicated by labels of the edges) has been indicated by the annotators. Among other tasks, this module adds certain points concerning coreference:

(i) the lemma of the node carrying the functor value ACT is assigned to the attribute COREF of an occurrence of the reflexive pronoun *se* that has not yet been treated (i.e. the PAT - Patient, Objective - of an active verb);

(ii) the remaining nodes without lemmas (in coordinated constructions or in apposition) are assigned the lemmas of their counterparts in the given construction; e.g. in *Jirka pozval Marii a Karel Milenu* (lit. 'Jirka invited Mary and Karel Milena'), the node corresponding to the deleted second occurrence of the verb (which has been added "by hand" as governing both *Karel*.ACT and *Milenu*.PAT) gets a lemma identical to that of the lefthand coordinated item.

The annotation on the underlying syntactic layer (the resulting structures being called **tectogrammatical** tree structures, TGTSs) is carried out in parallel in two streams both having as their inputs the result of the automatic preprocessing of the 'analytic' (surface) syntactic trees (in which every word token and every punctuation mark have their corresponding nodes and the basic kinds of dependency relations are specified); for a description of this procedure, see Böhmová and Hajičová (1999). The outputs of these streams differ in the size of data and the size of information carried by the tags:

(A) the set of "core" TGTSs (called 'large corpus', LC) has a large size, is being annotated with a higher speed and with tags carrying information about (a) the types of dependency relations and (b) values indicating the topic/focus articulation;

(B) the set of "full" TGTSs (the 'model' corpus, MC) has a smaller size, being annotated with a lower speed and with tags carrying complete tectogrammatical information (for a detailed characteristics of TGTSs, see Hajičová et al. 1999).

1.2. Since one of the aims of the PDT is to serve as a resource for linguistic research beyond the limits of the sentence, three specific attributes have been introduced in the TGTSs reflecting the **linking of sentences** to each other and to the context of situation:

(i) the attribute COREF having as its value the lexical value of the antecedent of the given anaphoric node (this node itself may be present on the surface, or deleted; the resolution of deletions is discussed by Hajičová and Sgall 2000),

(ii) the attribute CORNUM with a value equal to the serial number of the antecedent of the given node (to avoid uncertainty in case of two occurrences of the same word in the sentence), and (iii) the attribute CORSNT indicating whether the antecedent is in the same sentence (the value NIL) or in the preceding context (the value PREV). If an anaphoric node deleted on the surface is being restored, its lexical value is specified as an anaphoric (weak) pronoun (P in the sequel), a specific lexical value (L), or a technical value (such as Cor for the 'controllee').

1.3. The system of annotation of the TGTSs makes it possible to reflect the distinction between **grammatical** and **textual** coreference (see Panevová 1991). A typical example of the former is the coreference of the subject of the infinitival complementation of the control verbs (the subject gets the lexical value Cor) and the coreference of the reflexive pronouns (getting L identical to that of the subject), as well as that of the relative words in their relationship to their antecedents. With the latter kind of coreference (e.g. the 'deleted' pronominal subjects in Czech as a pro-drop language or other cases of pronominal reference) the nodes for the anaphoric expressions get P as their lexical value. Although also nouns, verbs, etc., can have a coreferential value, which we plan to reflect in the future shape of the procedure (in Czech, nouns in such a position often are accompanied by the pronoun (or determiner) *ten* 'that'), we do not discuss these cases in the present paper. In the case of grammatical coreference, the substantial feature of which is the presence of the antecedent in a specified syntactic position of the sentence, an additional attribute ANTEC is used with the value equal to the dependency relation (functor) of the antecedent.

2. TEXTUAL COREFERENCE

The textually coreferring node, which either corresponds to a **pronoun** or is a case of restored **deletion**, obtains a functor and a P lemma both in the MC and in the LC. In the MC, its attribute COREF obtains as its value the lemma of the antecedent, CORNUM gets the value of the serial

number of the antecedent (according to its word-order position, adjusted by decimal fractions in case of preceding deletion restorations); in CORSTN the unmarked value NIL is placed automatically, and changed into PREV if the antecedent is in the preceding sentence.

In the LC, the attribute COREF is left unfilled, and if the relevant node has been deleted, it is restored only in the case of a zero subject or of another deleted obligatory participant the head of which has not been deleted and is constituted by a deverbal noun or adjective of a fully productive type (as for deletion restoration, cf. Hajičová and Sgall, 2000; it should be noted that a restored node is always marked by the value ELID in one of its attributes).

In (1) and (2), we give examples of coreferential zero subjects in MC (we embrace the added nodes in square brackets):

(1) Udělal [on.ANIM.SG.ACT.ELID] to.

°He has done it°.

(2) Byla [ona.FEM.SG.PAT.ELID] předběhnuta několika jinými.

°She was left behind by some others°.

While with (1) the Gender value is based on intrasentential context (the properties of the verb), with (2) the clue is only present in intersentential context: *ona* is ambiguous (similarly as the forms *byla* and *předběhnuta*, on the base of the agreement with which it has been restored), having also the value 'they', NEUT.PL (e.g. if the neuter noun *děvčata* 'girls' is the antecedent). With most other pronominal forms the number will be supplied automatically, but Gender and the value of the Functor are filled in manually, which is necessary also in case the pronoun has not been deleted; only in certain specific cases an automatic solution is possible, e.g. with a plural noun in the Vocative case accompanying the subject, as in (3), or with the verb-subject agreement disclosing the Gender of the subject, as in (4):

(3) Vy jste, kluci, spali?

'You, boys, have been sleeping?'

Vy.ANIM.PL.ACT;COREF:kluk;CORNUM:4 jste, kluci, spali?

(4) My jsme tam byly všechny.

°We (women, girls) have been there all°.

My.FEM.PL.ACT jsme tam byly.FEM.PL všechny.

In (4), also some other attributes should be manually assigned their values if there is an antecedent in the previous sentence (otherwise just symbols for empty values are present). It may be recalled that a verb such as *prší* 'it rains' has no dependent ACT; its valency only admits adverbial adjuncts.

Under textual coreference also wider anaphoric relations are understood, which do not represent full referential identity, as e.g. in (5), in which *oni* 'they' is interpreted as referring to a group that includes Anna.

(5) Anna zase nepřišla. Oni všichni často chybějí. 'Anna failed to turn up again. They all often are absent.'

In the months to come, the automatic procedure is supposed to be enriched in various respects, to cover at least the most regular phenomena of several further subdomains, among which it is directly relevant for textual coreference that the development of the degrees of activation of the 'stock of shared knowledge' (see Hajičová 1993) will be registered as far as derivable from the use of nouns in subsequent utterances in a discourse.

3. GRAMMATICAL COREFERENCE

With grammatical coreference, the value of COREF is filled in (by the lemma of the controller, the subject or another antecedent, see below), along with the lemma of the coreferring node and with its functor, both in the LC and in the MC. In the MC, also the values CORNUM and ANTEC are added. In CORSTN, the unmarked value NIL remains, since with grammatical coreference the antecedent occurs in the same sentence.

The typical cases of grammatical coreference are reflexive and relative pronouns, and 'control':

3.1. **Reflexives**: With active clauses, in the second phase of the automatic procedure, the forms *se, si, sebe, sobě*, *sebou* (case and gender forms of 'himself') are assigned the value of COREF (i.e., the lemma of the subject); otherwise (with passive and with *svůj*, *svá*, etc., the possessive reflexive, which also has the lemma *se*) both the lemma and COREF are supplied manually.

The functor is determined on the basis of the values occurring in the 'analytic' structures; often the following syntactic values are concerned:

(i) *se* - PAT, ACT (general Actor, cf. (6)); in many cases the seemingly reflexive verb is not a true reflexive but just a lexical derivative, e.g. in *smát_se* 'laugh', *šířit_se* 'spread';

(6) To se má dělat rychle.

'One should do this quickly'

(ii) *si* - ADDR, PAT, BEN(efactive), or ETHD ('ethical dative'), e.g. in *dělejte si, co chcete* 'do whatever you wish';

(iii) *svůj* - se.APP, with Gender and Number of its antecedent.

A specific case is that of the reciprocal use of *se, si,* etc.; in the LC reciprocity is disregarded, but in the MC the pronoun gets the lemma *se*-Recp; most often the relation of reciprocity is constructed as coordination, and then it is the lemma of the conjunction that appears in COREF, see the example (7).

(7) Honza a Jirka se střídali.

'Johnny and George were alternating with each other.'

Honza.ACT a.CONJ Jirka.ACT se_Recp.PAT;COREF:a střídali.

Clauses with a plural subject are handled similarly, see (8).

(8) Chlapci se střídali.

'The boys were alternating with each other.'

Chlapci.ACT se_Recp.PAT;COREF:chlapec střídali.

In the MC, the attributes CORNUM and ANTEC get the values of the number and the functor of the antecedent, respectively.

3.2. **Relative** clauses are handled as congruent adjuncts of their antecedents; the functor of their verbs mostly is RSTR or DES (for restrictive and non-restrictive adjunct, respectively), both in LC and MC; the relative word gets its functor in accordance with its syntactic role within the clause, and the values of its attributes COREF, CORNUM in the MC correspond to the lemma and the number of the antecedent, as in (9) and (10).

(9) Jsou to lidé, kteří mají podobné názory.

°They are people who have similar opinions°

Jsou to lidé, kteří.ACT;COREF:lidé;CORNUM:3 mají.RSTR podobné názory.

(10) Jsme lidé, kteří se liší od zvířat...

°We are people, who differ from animals...'

Jsme lidé, kteří.ACT;COREF:lidé;CORNUM:2 se_liší.DES od zvířat...

Relative adverbs may have different functions, e.g. that of a Directional (as in (11)), not necessarily identical with that of the clause as a whole or with the anaphoric word accompanying it (and treated as its head).

(11) Kam to dáš, tam to najdeš.

'Where you put it there you find it.'

Kam.DIR-where_to;COREF:tam;CORNUM:4 to dáš.RSTR,
tam.LOC;COREF:kam;CORNUM:1 to najdeš.

3.3. The relation of **control** is handled manually for the time being, altough a part of the task is supposed soon to be fulfilled automatically. With most verbs of control the controller is specified as their Actor, Addressee or Patient. Due to the intrinsically syntactic character of the function of controller, we prefer to restore it in the form of a node labelled just with the 'technical' lemma Cor; in LC it gets the functor ACT (or, with a passive infinitive, PAT or ADDR) and with the lemma of the controller indicated in COREF. In MC also the functor of the controller and its position are filled in; see the following examples (with additions within LC again embraced in square brackets; the MC forms of (12) and (13) are (12') and (13'), respectively:

(12) Podnik plánoval [Cor.ACT.ELID;COREF:*podnik*] zvýšit výrobu.

°The firm planned to raise (its) production.°

(12') Podnik plánoval [Cor.ACT.ELID;COREF:*podnik*;CORNUM:1;

ANTEC:ACT] zvýšit výrobu.

(13) Radili synovi [Cor.ACT.ELID] k odchodu.

lit.: °They advised (their) son to departure.°

(13') Radili synovi [Cor.ACT.ELID;COREF:*syn*;CORNUM:2;

ANTEC:ADDR] k odchodu.

Note: We distinguish between *Jirka slíbil přijít* °George promised to come°, where a node with the lemma *Cor* functions as ACT of the infinitive (since the alternative that someone else would be coming is out of question) and a structure with textual coreference as e.g. *Jirka slíbil, že přijde* °G. promised that he would come°, where as ACT of the infinitive the personal pronoun *on* with gender assigned according to context is supplied (in this case the alternative that but someone else would be coming rather than Jirka is quite possible).

As (13) shows, also nouns of action functioning as objects of a verb of control are treated in this way. This concerns also the so-called Slavonic infinitive with accusative (the verb *slyšet* 'hear' has the frame ACT PAT (EFF) on this reading, i.e. the 'second object', Effect, is optional); (14) and (14') illustrate the assignment of tags in LC and MC, respectively.

(14) Honza slyšel Karla [Cor.ACT.ELID;COREF:Karel] otvírat dveře.

°Johnny heard Charles open the door°

(14') Honza slyšel Karla [Cor.ACT.ELID CORNUM:3;ANTEC:PAT] otvírat dveře.

If the position of PAT is not occupied by a specific lexeme in this construction, as in (15), then the lexical value in the COREF attribute is Gen (denoting a general participant):

(15) Jan slyšel [Cor.ACT.ELID;COREF:Gen] otvírat dveře.

°John heard the door open.°

4. CONCLUDING REMARKS

By now, 100 000 sentences from the Czech National Corpus have obtained their 'analytic' annotations, and we expect to get several thousands of sentences annotated by their TGTSs before the end of the year 2000.

Neither the automatic nor the manual part of the tagging can achieve a complete formulation of tectogrammatical representations. Several types of grammatical information will be specified only after further empirical investigations. Thus, e.g., the disambiguation of the functions of prepositions and conjunctions can only be completed after lists of nouns and verbs with specific syntactic properties are established. However, the annotated corpus will offer a suitable starting

point for monographic analysis of the problems concerned. Whenever possible, also statistical methods will be used.

In this way a theoretically substantiated labelling of the TRs can be gained, distinguishing between different kinds of objects and adverbials, between meanings of function morphemes, topic and focus, and so on. The result will be much more complex than that of a parser or tagger of the usual kinds: not only the grammatical well-formedness will be checked, but disambiguated representations of sentences will be achieved, which (although underspecified in the points in which the sentence structure is not fully specific - indistinctness, "systematic ambiguity", scopes of quantifiers) would constitute an appropriate input for a procedure of semantic(-pragmatic) interpretation.

Acknowledgement:

The research reported on in this paper has been predominantly carried out within the project supported by the Czech Grant Agency 405-96-K214, and in part by that of the Czech Ministry of Education VS 96-151.

References:

Böhmová A. and E. Hajičová (1999). The Prague Dependency Tree Bank I: How much of the underlying syntactic structure can be tagged automatically? *The Prague Bulletin of Mathematical Linguistics* 71, 5-12.

Hajič J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In: *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, ed. by E. Hajičová, 106-132. Prague: Karolinum.

Hajičová E. (1993). *Issues of sentence structure and discourse patterns.* Prague: Charles University.

Hajičová E. (1998). Prague Dependency Treebank: From analytic to tectogrammatical annotations. In: *Text, Speech, Dialogue*, ed. by P. Sojka, V. Matoušek, K. Pala and I. Kopeček), Brno: Masarykova univerzita, 45-50.

Hajičová E., J. Panevová and P. Sgall (1999). *Manuál pro tektogramatické značkování Českého národního korpusu* [Manual for tectogrammatical tagging of the Czech National Corpus], Tech. Report No. 7, Institute of Formal and Applied Linguistics, Charles University, Prague. English translation to be published in 2000.

Hajičová E. and P. Sgall (2000). Semantico-syntactic tagging of very large corpora: The case of restoration of nodes on the underlying level. In this volume.

Panevová J. (1991). Koreference gramatická nebo textová? [Grammatical or textual coreference?] In: *Etudes de lingustique romane et slave*, ed. by W. Banys, L. Bednarczuk, and K. Bogacki), Cracow, 495-506.