Semantico-Syntactic Tagging of Very Large Corpora: the Case of Restoration of Nodes on the Underlying Level

Eva Hajičová and Petr Sgall

Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25, 11800 Praha 1, Czechia
{hajicova, sgall}@ufal.mff.cuni.cz

Abstract

The Prague Dependency Treebank has been conceived of as a semi-automatic three-layer annotation system, in which the layers of morphemic and 'analytic' (surface-syntactic) tagging are followed by the layer of tectogrammatical tree structures. Two types of deletions are recognized: (i) those licensed by the grammatical properties of the given sentence, and (ii) those possible only if the preceding context exhibits certain specific properties. Within group (i), either the position itself in the sentence structure is determined, but its lexical setting is 'free' (as e.g. with a deleted subject in Czech as a pro-drop language), or both the position and its 'filler' are determined. Group (ii) reflects the typological differences between English and Czech; the rich morphemics of the latter is more favorable for deletions. Several steps of the tagging procedure are carried out automatically, but most parts of the restoration of deleted nodes still have to be done "manually". If along with the node that is being restored, also nodes depending on it are deleted, then these are restored only if they function as arguments or obligatory adjuncts. The large set of annotated utterances will make it possible to check and amend the present results, also with applications of statistic methods. Theoretical linguistics will be enabled to check its descriptive framework; the degree of automation of the procedure will then be raised, and the treebank will be useful for most different tasks in language processing.

1. Introductory remarks

The large corpus built in the Institute of Czech National Corpus (led by F. Čermák) at the Faculty of Philosophy, Charles University, Prague, now comprises more than 100 millions of word occurrences from different kinds of texts. A part of this corpus has been used as the basis of the Prague Dependency Treebank (PDT, see Hajič 1998), the scenario of which is based on the conviction of the initiators of the project that the result of tagging is to be used both for the purposes of empirical and theoretical linguistic research and for its 'practical' applications, such as in dictionary making or in the build-up of different systems of natural language processing. The PDT is therefore conceived of as a semi-automatic three-layer annotation system (see Hajičová 1998), in which the layers of morphemic and 'analytic' (surface-syntactic) tagging are followed by a third layer, viz. that of tectogrammatical tree structures (TGTSs in the sequel). The TGTSs are intended to represent the underlying syntactic structure of sentences, which would be appropriate as the input to semantic(-pragmatic) interpretation, since the irregularities of the shallow layers, including synonymy and ambiguity, are absent on this level (see Hajičová 1993 and the writings quoted there). This implies that in the TGTSs, nodes for cases of (surface) deletions should be added ('reconstructed').

The transition from morphemic and analytic to tectogrammatical annotations is divided into three steps:

first, an automatic procedure changes the morphemic tags into the corresponding grammatemes (values of morphological categories: tense, number, etc.), whenever possible, combining every analytic word form into a single node (the label of which contains the lexical lemma indexed with a string of grammatemes derived from endings and grammatical affixes, as well as from auxiliary verbs, articles, prepositions, conjunctions; the only exceptions are coordinating conjunctions, which retain their nodes as governors of the coordinated syntagm);

second, a manual step is devoted to specify most of the syntactic relations (functors) and the more difficult cases of grammatemes (including those reflecting the topic-focus articulation of the sentence and corresponding movements);

third, another automatic step takes care for specifications that can be carried out on the basis of the preceding step, i.e. after the syntactic functions of the lexical occurrences have been fully determined (cf. e.g. the values of the pronouns discussed under ex. (5) below).

By now, 100 000 sentences from the Czech National Corpus have obtained their analytic annotations, and we expect to get thousands of sentences annotated by their TRs before the end of the year 2000. Hundreds of sentences (the 'large corpus', LC, have already been tectogrammatically tagged as for the main points, including the restoration of most of the deleted items. A more detailed annotation has been achieved, up to now, for about 100 sentences (the 'model corpus', MC).

2. Types of deletion

Our preliminary analysis of the Czech National Corpus indicates that the following types of deletions have to be recognized:

(i) deletions licensed by the grammatical properties of sentence elements or sentence structure,

(ii) deletions possible only if the preceding context (be it co-text or context of situation) exhibits certain specific properties.

Our subclassification of reconstructions of nodes can be compared with the kinds of 'silent' anaphora in the annotation scheme of the FrameNet project (Fillmore 1999); in the latter, it is especially the case (b) in Section 3 below (that of a "zero morph") the counterpart of which has been elaborated in detail. We do not go so far e.g. in the analysis of deverbative nouns, i.e. we just exclude the Actor from the valency frame of an agentive noun, such as *writer*, instead of characterizing the suffix as filling this slot. Other participants of verbs and deverbative nouns are restored, even if their head itself has been deleted and has to be added. In the LC, we in principle do not restore any (deleted) complementations of nouns except for the case of the maximally productive deverbatives with the

protypical suffix *-á/aní, -tí* (eg. *čekání* 'waiting' from the verb *čekat* 'to wait'; we distinguish between *psaní* 'a letter' and 'writing' as in *Dostali jsme psaní* 'We got a letter' and *Psaní mu trvalo hodinu* 'Writing took him an hour').

3. Grammatical identification of the deleted item

Within group (i), two situations may obtain:

(a) Only the position itself in the sentence structure is predetermined (i.e. a sentence element is subcategorized for this position), but its lexical setting is 'free'. This is e.g. the case given by the so-called pro-drop character of a language like Czech, where the position of the subject of a verb is 'given', but it may be filled in dependence on the context, cf. (1):

(1) Předseda vlády řekl, že předloží návrh na změnu volebního systému.
'The Prime-minister said that (he - the Prime-minister, the Government, or somebody else identifiable on the basis of the context) will submit a proposal on the change of the election system').

Here also belong cases of the semantically obligatory but deletable complementations of verbs: e.g. the Cz. verb *přijet* 'to arrive' has as its obligatory complementation an Actor and a Directional "where-to" (the obligatoriness of the Directional complementation can be tested by a question test, see Panevová 1974; Sgall et al. 1986), which can be deleted on the surface, cf. (2); here the Directional (*here* or *there*) is deleted because the speaker assumes that the hearer will identify the referent easily).

(2) Vlak přijede v šest hodin.
'The train will arrive at six o'clock.'

Also a subject to a verb is supplied if it fails to be expressed in the surface, Cz. being a pro-drop language. A node with a label containing the lemma of a personal pronoun (including the anaphoric 3rd person pronoun) is added, and its values of gender and number are specified according to the congruent form of the verb and to what has been understood from the intra- or intersentential context; the restered node obtains also a functor (ACT - Actor, Dir-3 - Directional 'where to'). In the following examples, the added values are inserted in square brackets; a restored node always is marked as deleted (elided) by the value ELID:

(3) [My.ANIM.PL.ACT.ELID] Byli jsme tam všichni.
'We all were there.'

(4) Marie a Jana [tam.DIR-3.ELID] přišly a [on.FEM.PL.ACT.ELID] posadily se na pohovku;
'Mary and Jane came [there] and [they] sat down on the sofa.

(5) Děti rozbily okno, ale [on.FEM.PL.ACT.ELID] omluvily se.
The children broke a window, but [they] apologized.

The value of Number with 1st pers. pronouns will be supplied both in MC and in LC by the second phase of the automatic procedure, on the input of which the subject-verb agreement has been specified. With the 3rd pers. pronouns, in MC also the functor of antecedent and its serial number in the word order will be marked as values of specific attributes (distinguishing whether the antecedent occurs in the given sentence or in its predecessor in the text).

The supplied word is always placed to the left of its governing word (should more of them be inserted into one and the same place, then it must conform to the systemic ordering, i.e., ACT followed by most of the free modifications, then ADDR, PAT and EFF in this order).

(b) Both the position and its 'filler' is predetermined; this situation might be described as the presence of a "zero morph" rather than deletability, especially in case the deletion is obligatory. An example of the function of a zero morph are the so-called General Participants:

(6) Ta kniha [Gen.ACT.ELID] byla už vydána dvakrát.
'The book has already been published twice.' Actor)

(7) V neděli [Gen.PAT.ELID] obvykle peču.
'lit.: On Sundays (I) usually bake'

(8) Dědeček [Gen.ADDR.ELID] vypravuje pohádky.
'Grandfather tells fairytales'

Also the phenomena of 'control' belong here, see Hajičová, Panevová and Sgall (2000).

4. Contextual identification of the deleted item

Group (ii) consists of deletion conditioned by the context, with which the item to be restored is determined by the context alone. This is a point where the typological differences between English and a language with rich inflection, such as Czech, are most clearly to be seen; the rich morphemics allows for deletions in many cases in which a deletion is impossible in the English text. To put it in an extreme way, in principle everything in any position can be deleted in Czech if it is identifiable on the basis of the context; this is not the case in English, cf. e.g. the deletion of the whole topic of the sentence in (9):

(9) (Potkal jsi včera Toma?) Potkal.
'lit.: (Did you meet Tom yesterday?) Met'.

Along with these rather specific cases (in which the verb in a typical context does require the Objective to be present also in the surface), two cases are characteristic of contextual deletion in Czech:

(a) The restored node (i.e. deleted in the surface) is a governor of a congruent adjective which has the functor ExD in the manually prepared analytic trees):

(10) Přišli jen [ten.ACT.Plur.ELID] mladší.
°Only (the) younger [ones] came.'

(11) Našli jen [ten.PAT.Plur.ELID] modré.
°They only found (the) blue (ones).°

This doesn°t concern those adjectival words with which we assume a substantival function as well: such pronouns as *ten (to)* 'this', *některý* 'some', cardinal numerals, superlatives, and of course the 'substantivized adjectives', (*nemocný*, °ill', *raněný* 'wounded°, etc.):

(12) Zvolili tři.PAT z pěti místopředsedů.
°They elected three from the five vice-presidents°

(13) (Připravili večeři pro deset hostů.) Přišli jen čtyři.ACT

A noun is not restored with adjectival words in constructions with the functor PAT with a copula (*Kluci*

*byli úspěšní.*PAT °the boys were successful°) or with EFF and COMPL: e.g., *pokládat za své.*EFF °regard as (one°s) own°, *našli je nemocné.*COMPL °they found them ill°.

(b) In coordination structures nodes for the deleted repetitions of the governing word are restored in certain cases.

Often the deleted word depends directly on the node COORD. However, we prefer to choose the simpler structure with a single lexical head as long as this is not excluded by clear semantic or syntactic factors. Thus, (14) is handled as not including deletion, since the two coordinated predicates can be understood as to be 'jointly' modified by the two arguments.

(14) Jirka potkal a pozdravil Marii.
°George met and greeted Mary°

Sometimes adding a node is inevitable: in (15) the presence of deletion is clearly given by the fact that some of the dependents of the two heads (one of which is deleted) differ:

(15) Potkal Marii včera a já [jsem Marii potkal] dnes.
°He met Mary yesterday and I [met Mary] today.°

In such a case, the verb is restored manually, but its Objective is then specified by the second part of the automatic procedure, which identifies it in accordance with the lefthand branch of the coordination construction. The lemma of the new node gets the shape of a noun, rather than a pronoun, since the presence of coreferentiality is necessary in such a sentence, which would not be the case with e.g. (16), in which the pronoun *ji* 'her' can refer to another person, if this has been strongly activated by the context:

(16) Potkal Marii včera a já jsem ji potkal dnes.
°He met Mary yesterday and I met her today.°

In other cases the inevitability of restoring a deleted item is given semantically, cf. (17), in which the two adjuncts cannot be interpreted as depending on a single occurrence of the noun:

(17) Pil červené a bílé víno.
°He drank red and white wine°.

We are aware of the difficulties connected with drawing such a boundary between sentences with and without deletion. On the one hand, the theory of language cannot exclude the possibility that once a kind (or way of existence) of wine comes about that would somehow adopt two colors, cf. e.g. such noun groups as *a red and white flag*. On the other hand, there are cases with which the annotators have to look for clues in a broader context, and perhaps do not find them (cf. the much discussed example of *old men and women*). However, as far as practical issues of natural language processing are concerned, the present preliminary solution seems to be relatively suitable, at least before very large sets of examples can be studied (which will only be possible on the basis of very large syntactically tagged corpora). It does not seem to be crucial that some of the cases under (b) also meet the conditions of (a) above.

5. Concluding remarks

If along with the node that is being restored, also one or more nodes depending on it are deleted, then in LC these are restored only if they function as arguments (rather than adjuncts) of their head; adjuncts ('free complementations') are concerned only if they are obligatory with the given head. In MC the symbol ELEX is distinguished from ELID: ELEX is assigned to the restored node when it is necessary - according to the meaning of the given sentence - to add its optional adjuncts, which, however, we do not specify; e.g., with (18) the verb is restored, but the optional adjunct *včera* 'yesterday' is not restored in the second part of the conjunction; the existence of a further modification that could be transferred from the first clause gets merely indicated by adding *navštívit.*CO.ELEX as the rightmost member of the coordination structure);

(18) Včera navštívil Jirka Marii a Milan Jiřinu.
°Yesterday George visited Mary and Milan Georgine.°

Wherever the lexical unit can be restored in a non-systemic way, yet univocally (it may be from the preceding sentences, crossing the full stop), in MC the lemma of the restored unit is added to the node with the supplied pronominal element, but only as the value of the grammateme COREF: *Přišel k Jiřině a dal jí kytku* °He came to J. and gave her a posy° - the pronominal lemma *jí* will be assigned: on.FEM.SG.ADDR, and, if such an utterance clearly asserts that the person he gave a bunch of flowers was not someone else, whose image has been activated by the context, the value of the attribute COREF will be Jiřina, and in CORNUM the serial number of *Jiřině* will be registered.

The brief characterization of how the deleted nodes are restored in PDT could only illustrate some aspects of the first steps of the procedure of semi-automatic syntactic tagging. For most further steps it will be possible to use (also with applications of statistic methods) the large set of annotated utterances obtained in this way. The existence of such a set will make it possible to check the errors present there on the basis of monographic studies, which will find a much more suitable starting point with the syntactically tagged corpus than with the previously used excerpts.

With a large corpus tagged not only on the level of morphemics, but also on that of (underlying) syntax, theoretical linguistics will be enabled to check its descriptive framework; this is quite a new situation for our branch of science.

Moreover, such a treebank will help rise not only the degree of precision of the analysis, but also that of its automation. The usefulness of such a corpus for most different tasks in language processing - from information mining and machine (assisted) translation to communication with intelligent systems - will also be strengthened.

6. 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.

7. References

Fillmore, Charles J. (1999). Silent anaphora. Paper presented at the TELRI Workshop, November 1999, Bratislava.

Hajič, Jan (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, Eva (1993). *Issues of sentence structure and discourse patterns.* Prague: Charles University.

Hajičová, Eva (1998). Prague Dependency Treebank: From analytic to tectogrammatical annotations. In: *Text, Speech, Dialogue* (eds. Petr Sojka, Václav Matoušek, Karel Pala and Ivan Kopeček). Brno: Masarykova univerzita, 45–50.

Hajičová, Eva, Jarmila Panevová, Petr Sgall (2000). Coreference in annotating large corpora. In this volume.

Panevová, Jarmila (1974). On verbal frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics.* 22:3–40, 23(1975):17–52; a revised version in *Prague Studies in Mathematical Linguistics* 6 (1978), 227–254.

Sgall, Petr, Eva Hajičová, Jarmila Panevová (1986). *The meaning of the sentence in its semantic and pragmatic aspects.* Ed. by J. L. Mey. Dordrecht: Reidel / Prague: Academia.