Deletions and their reconstruction in tectogrammatical syntactic tagging of very large corpora

Eva HAJIČOVÁ
ÚFAL
Charles University
Malostranské nám. 25
118 00 Prague, Czech Republic
hajicova@ufal.mff.cuni.cz

Markéta CEPLOVÁ
ÚFAL
Charles University
Malostranské nám. 25
118 00 Prague, Czech Republic
ceplovam@yahoo.com

Abstract

The procedure of reconstruction of the underlying structure of sentences (in the process of tagging a very large corpus of Czech) is described, with a special attention paid to the conditions under which the reconstruction of ellipted nodes is carried out.

1. The tagging scenarios with different (degrees and types of) theoretical backgrounds have undergone a rather rapid development from morphologically based part-of-speech (POS) tagging through treebanks capturing the surface syntactic structures of sentences to semantically oriented tagging models, taking into account the underlying structure of sentences and/or certain issues of the 'inner' semantics of lexical units and their collocations.

One of the critical aspects of the tagging scenario capturing the underlying structure of the sentences is the 'depth' of the resulting tree structures; in other words, how far these structures differ from the surface structures. If we take for granted (as is the case in most of the syntactic treebanks) that every word of the (surface) sentence should have a node of its own in the surface tree structure, then this issue can in part be reformulated in terms of two subquestions:

(i) which surface nodes are superfluous and should be 'pruned away',

(ii) which nodes should be assumed to be deleted in the surface and should be 'restored' in the underlying structure (e.g. in forms of different kinds of dummy symbols, see Fillmore 1999).

In our paper, we are concerned with the point (ii).

2. In the TG and post-TG writings, it is common to distinguish between two types of deletions: (a) ellipsis proper and (b) gapping. For both of them, it is crucial that the elliptical construction and its antecedent should be parallel and 'identical' at least in some features. The two types of ellipsis can be illustrated by examples (1) and (2), respectively.

(1) Psal jenom úkoly, které chtěl.
lit. 'He-wrote only homework's which he-wanted'

(2) Honza dal Marii růži a Petr Idě tulipán.
lit. 'John gave Mary rose and Peter Ida tulip'

For both types, a reconstruction in some way or another is necessary, if the tree structure is to capture the underlying structure of the sentences.

3. The examples quoted in the previous section cover what Quirk et al. (1973, pp. 536-620) call 'ellipsis in the strict sense'; they view ellipsis as a purely surface

phenomenon: the recoverability of the ellipted words is always unique and 'fits' into the surface structure. They differentiate ellipsis from 'semantic implication' which would cover e.g. such cases as (3) and (4):

(3) John wants to read.

(4) Thanks.

If (3) is 'reconstructed' as 'John wants John to read', then the two occurrences of 'John' are referentially different, which is not true about the interpretation of (3). With (4), it cannot be uniquely determined whether the full corresponding structure should be 'I owe you thanks' or 'I give you thanks' etc.

4. For tagging a corpus on the underlying level, it is clear that we cannot limit ourselves to the cases of ellipsis in the strict sense but we have to broaden the notion of 'reconstruction' to cover both

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

(ii) deletions licensed only by the preceding context (be it co-text or context of situation).

4.1. In our analysis of a sample of Czech National Corpus, two situations may occur within the group (i):

(a) Only the position itself that should be "filled" 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 of the so-called pro-drop character of Czech, where the position of the subject of a verb is 'given', but it may be filled in dependence on the context.

(5) Předseda vlády řekl, že předloží návrh na změnu volebního systému.
'The Prime-minister said that (0) will submit a proposal on the change of the electoral system.'

The 'dropped' subject of the verb *předloží* 'will submit' may refer to the Prime-minister, to the Government, or to somebody else identifiable on the basis of the context.

Here also belong cases of the semantically obligatory but deletable complementations of verbs: the Czech 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; its reference is determined by the context.

(6) Vlak přijede v poledne.
'The train will arrive at noon.'

The utterer of (6) deletes the Direction 'where-to' because s/he assumes that the hearer knows the referent.

(b) Both the position and its 'filler' are predetermined.

This is the case of e.g. the subject of the infinitival complement of the so-called verbs of control as in (7).

(7) Předseda vlády slíbil předložit návrh na změnu volebního systému.
'The Prime-minister promised to submit a proposal on the change of the electoral system.'

The identification of the underlying subject of the infinitive is 'controlled' by the Actor of the main verb, in our example it is 'the Prime-minister'.

Another example of this class of deletions are the so-called General Participants (close to the English *one* or German *man*): General Actor in (8), General Patient in (9), or General Addressee in (10).

(8) Ta kniha byla už vydána dvakrát.
'The book has already been published twice.'

(9) V neděli obvykle peču.
'On Sundays (I) usually bake.'

(10) Dědeček často vypravuje pohádky.
'Grandfather often tells fairy-tales.'

4.2 Within the group (ii), there belong cases of the so-called 'occasional ellipsis' conditioned by the context alone.

We are aware that not everything in any position that is identifiable on the basis of the context can be deleted in Czech (as might be in an extreme way concluded from examples (11) through (14)). However, the conditions restricting the possibility of ellipsis in Czech seem to be less strict than e.g. in English, as illustrated by (15):

(11) Milujeme a ctíme svého učitele.
'We love and honour our teacher.'

(12) Marii jsem viděl a slyšel zpívat.
lit. 'Mary-Acc. Aux-be saw and heard to-sing'
'I saw and heard Mary singing.'

(13) Jirka se včera v hospodě opil do němoty a Honza dneska.
lit. 'Jirka himself yesterday in pub drunk to death and Honza today.'
'In the pub, Jirka drunk himself to death yesterday and Honza today.'

(14) Petr říkal Pavlovi, aby šel ven, a Martin, aby zůstal doma.
'Peter told Pavel to go outside and Martin (told Pavel) to stay at home.'

(15) (Potkal jsi včera Toma?) Potkal.
'(Did you meet Tom yesterday?) Met'.

4.3 In addition to setting principles of which nodes need to be restored it is also important to say in which cases no restoration is desirable. Nodes are not restored in cases of:

(a) accidental omission (due to emotion, excitement or insufficient command of language, see e.g. Hlavsa 1990);

(b) unfinished sentences, which usually lack focus (unlike ellipsis where the 'missing' elements belong to topic);

(c) sentences without a finite verb that can be captured by a structure with a noun in its root (in these cases there are no empty positions, nothing can be really added).

All these cases have no clear-cut boundaries, rather it is more appropriate to expect continual transitions.

5.1 The Prague Dependency Tree Bank (PDT in the sequel), which has been inspired by the build-up of the Penn Treebank (Marcus, Santorini & Marcinkiewicz 1993; Marcus, Kim, Marcinkiewicz et al. 1994), is aimed at a complex annotation of (a part of) the Czech National Corpus (CNC in the sequel), the creation of which is under progress at the Department of Czech National Corpus at the Faculty of Philosophy, Charles University (the corpus currently comprises about 100 million tokens of word forms). PDT comprises three layers of annotations: (i) the morphemic layer with about 3000 morphemic tag values; a tag is assigned to each word form of a sentence in the corpus and the process of tagging is based on stochastic procedures described by Hajič and Hladká (1997); (ii) analytic tree structures (ATSs) with every word form and punctuation mark explicitly represented as a node of a rooted tree, with no additional nodes added (except for the root of the tree of every sentence) and with the edges of the tree corresponding to (surface) dependency relations; (iii) tectogrammatical tree structures (TGTSs) corresponding to the underlying sentence representations, again dependency-based.

At present the PDT contains 100000 sentences (i.e. ATSs) tagged on the first two layers. As for the third layer, the input for the tagging procedure are the ATSs; this procedure is in its starting phase and is divided into (i) automatic preprocessing (see Böhmová and Sgall 2000) and (ii) the manual phase. The restoration of the syntactic information absent in the surface (morphemic) shape of the sentence (i.e. for which there are no nodes on the analytic level) is mostly (but not exclusively) done –

at least for the time being – in the manual phase of the transduction procedure. In this phase, the tagging of the topic-focus articulation is also performed (see Buráňová, Hajičová and Sgall 2000).

5.2 The reconstruction of deletions in TGTSs is guided by the following general principles:

(i) All 'restored' nodes standing for elements deleted in the surface structure of the sentence but present in its underlying structure get marked by one of the following values in the attribute DEL:

ELID: the 'restored' element stands alone; e.g. the linearized TGTS (disregarding other than structural relations) for (16) is (16'). (Note: Every dependent item is enclosed in a pair of parenthesis. The capitalized abbreviations stand for dependency relations and are self-explaining; in our examples we use English lexical units to make the representations more transparent.)

(16) Sbíral houby.
'Collected-he mushrooms.'

(16') (he.ACT.ELID) collected (mushrooms.PAT)

ELEX: if the antecedent is an expanded head node and not all the deleted nodes belong to the obligatory complementations of the given node and as such not all are reconstructed, cf. e.g. the simplified TGTS for (13) in (13').

(13') ((Jirka.ACT) (yesterday.TWHEN) (pub.LOC) drunk-himself (to-death.MANN)) and (drunk-himself.ELEX (Honza.ACT) (today.TWHEN))

EXPN: if the given node itself was not ellipted but some of its complementations were and are not restored (see the principle (iii)(b) below), cf. e.g. the simplified TGTS in (15') for (15) above, with non-reconstructed temporal modification:

(15') (I.ACT.ELID) met.EXPN (Tom.PAT.ELID)

(ii) The restored nodes are added immediately to the left of their governor.

(iii) The following cases are prototypical examples of restorations (for an easier reference to the above discussion of the types of deletions, the primed numbers of the TGTSs refer to the example sentences in Section 4):

(a) Restoration of nodes for complementations for which the head nodes (governors) are subcategorized. The assignment of the lexical labels is governed by the following principles: in pro-drop cases (5') (comparable to Fillmore's 1999 CNI – constructionally-licensed null instantiation) and with an obligatory but deletable complementation (6') (cf. Fillmore's definite null instantiation, DNI) the lexical value corresponds to the respective pronoun; with grammatical coreference (control), the lexical value is Cor (7'); in both these cases, the lexical value of the antecedent is put into a special attribute of Coreference; in cases of general participants (cf. Fillmore's indefinite null instantiation – INI) the lexical value is Gen (10'):

(5') (prime-minister.ACT) said ((he.ACT.ELID; COREF: prime-minister) will-submit.PAT (proposal.PAT (change.PAT (system.PAT (electoral.RSTR)))))

(6') (train.ACT) will-arrive (noon.TWHEN) (here/there.ELID.DIR3)

(7') (prime-minister.ACT) promised ((Cor.ACT.ELID; COREF: prime-minister) submit.PAT (proposal.PAT (change.PAT (system.PAT (electoral.RSTR)))))

(10') (grandfather.ACT) (often.TWHEN) (Gen.ADDR.ELID) tells (fairy-tales.PAT)

(b) Elipted optional complementations are not restored (see (13') above) unless they are governors of adjuncts.

(c) For coordinated structures, the guiding principle says: whenever possible, give precedence to a "constituent" coordination before a "sentential" one (more generally: "be as economical as possible"), thus examples like (17) are not treated as sentential coordination (i.e. they are not transformed into structures corresponding to (17')).

(17) Karel přinesl Janě květiny a knihu.
'Karel brought Jane flowers and a book.'

(17') Karel přinesl Janě květiny a Karel přinesl Janě knihu.
'Karel brought Jane flowers and Karel brought Jane a book.'

A special symbol CO is introduced in the complex labels for the coordinated nodes to mark which nodes stand in the coordination relation and which modify the coordination as a whole (see (11')); the lexical value of the restored elements is copied from the antecedents (see (13') above):

(11') ((we.ACT) (love.CO) and (honour.CO) (our teacher.PAT))

The analysis of (11') is to be preferred to sentential coordination with deletion also for its correspondence with the fact, that in Czech object can stand after coordinated verbs only if the semantic relation between the verbs allows for a unifying interpretation, as shown by cases, where the object must be repeated with each verb (compare the contrast between (18) and (19)).

(18) Potkal jsem Petra, ale nepoznal jsem ho.
'I met Peter, but I didn't recognize him.'

(19) ??Potkal, ale nepoznal jsem Petra.
'I met but didn't recognize Peter.'

However, there are cases where the coordination has to be taken as sentential or at least at a higher level. As modal verbs are represented as gramatemes of the main verb, sentences as (20) have to be analysed as in (20'):

(20) Petr musel i chtěl přijít.
'Peter had to and wanted to come.'

(20') (Peter.ACT) (had-to-come.CO) and (wanted-to-come.ELID. CO)

Another case of a less strict adherence to the economy principle are sentences with double reading. Such a treatment then allows for a distinction to be made between the two readings, e.g. in (21), namely between (a) 'villagers who are (both) old and sick' and (b) 'villagers who are sick (but not necessarily old) and villagers who are old (but not necessarily sick)':

(21) Jim zachránil staré a nemocné vesničany.
'Jim saved old and sick villagers.'

(21'a) (Jim.ACT) saved (villagers.PAT ((old.CO.RSTR) and (sick.CO.RSTR)))

(21'b) (Jim.ACT) saved ((villagers.CO.PAT.ELID (old.RSTR)) and (villagers.CO.PAT (sick.RSTR)))

5.3 The research reported on in this contribution is work in progress: the principles are set, but precisions are achieved as the annotators progress. There are many issues left for further investigation; let us mention just one of them, as an illustration. Both in (22) and in (23), the scope of 'málokdo' (few) is (at least on the preferential readings) wide ('there are few people such that...'); however, (24) is ambiguous: (i) there were few people such that gave P. a book and M. flowers, (ii) few people gave P. a book and few people gave M. flowers (not necessarily the same people). A similar ambiguity is exhibited by (25): (i) there was no such (single) person that would give P. a book and M. flowers, (ii) P. did not get a book and M. did not get flowers. However, there is no such ambiguity in (26).

(22) Málokdo jí jablka a nejí banány.
lit. 'Few eat apples and do-not-eat bananas'
'Few people eat apples and do not eat bananas.'

(23) Málokdo dal Petrovi knihu a Marii květiny ne.
lit. 'Few gave Peter book and Mary flowers not'
'Few people gave Peter a book and did not give Mary flowers.'

(24) Málokdo dal Petrovi knihu a Marii květiny.
lit. 'Few gave Peter book and Mary flowers'
'Few people gave Peter a book and Mary flowers.'

(25) Nikdo nedal Petrovi knihu a Marii květiny.
lit. 'Nobody did-not-give Peter book and Mary flowers'
'Nobody gave Peter a book and Mary flowers.'

(26) Petrovi nikdo nedal knihu a Marii květiny.
lit. 'Peter nobody did-not-give book and Mary flowers'
'To Peter, nobody gave a book and to Mary, flowers.'

An explanation of this behaviour offers itself in terms of the interplay of contrast in polarity and of topic-focus articulation: an element standing at the beginning of the sentence with a contrast in polarity carries a wide scope ('few' in (22) and (23)); with sentences without such a contrast both wide scope and narrow scope interpretations are possible ('few' and 'nobody' in (24) and (25), respectively); (25) differs from (26) in that in the latter sentence, the element in contrastive topic is 'Peter' in the first conjunct and 'Mary' in the second, rather than 'nobody', and there is no contrast in polarity involved.

The tagging scheme sketched in the previous sections offers only a single TGTS for the ambiguous structures instead of two, which is an undesirable result. However, if the explanation offered above is confronted with a larger amount of data and confirmed, the difference between the two interpretations could be captured either by means of a combination of tags for the restored nodes and for the topic-focus articulation or by different structures for coordination: while example (22) supports the economical treatment of coordinate structures (the ACT modifying the coordination as whole), examples (24) through (26) seem to suggest that there may be cases where the other approach (sentential coordination with ellipsis) is more appropriate to capture the differences in meaning.

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

References

Böhmová A. and Sgall P. (2000) *Automatic procedures in tectogrammatical tagging*. In these Proceedings.

Buráňová E., Hajičová E. and Sgall P. (2000) *Tagging of very large corpora: Topic-Focus Articulation*. In this volume.

Fillmore C. J. (1999) *Silent anaphora: Corpus, FrameNet, and missing complements*. Paper presented at the TELRI workshop, Bratislava.

Hajič J. and Hladká B. (1997) *Probabilistic and rule-based tagger of an inflective language – a comparison*. In "Proceedings of the Fifth Conference on Applied Natural Language Processing", Washington, D.C., pp. 111-118.

Hajičová E., Panevová J. and Sgall P. (1998) *Language Resources Need Annotations To Make Them Really Reusable: The Prague Dependency Treebank*. In "Proceedings of the First International Conference on Language Resources & Evaluation", Granada, Spain, pp. 713-718.

Hajičová E., Partee B. and Sgall P. (1998) *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer, Dordrecht.

Hlavsa Z. (1990) *Some Notes on Ellipsis in Czech Language and Linguistics*. Studi italiani di linguistica teorica ed applicata 19, pp. 377-387.

Marcus M. P., Kim G., Marcinkiewicz M. A. et al. (1994) *The Penn Treebank: Annotating Predicate*

Argument Structure. Proceedings of the ARPA Human Language Technology Workshop. Morgan Kaufmann, San Francisco.

Marcus M. P., Santorini B. and Marcinkiewicz M. A. (1993) *Building a Large Annotated Corpus of English: the Penn Treebank*. Computational Linguistics, 19(2), pp. 313-330.

Panevová J. (1974) *On verbal frames in Functional Generative Description*. Prague Bulletin of Mathematical Linguistics 22, pp. 3-40; 23(1975), pp. 17-52.

Quirk R., Greenbaum S., Leech G. and Svartvik J. (1973) *A grammar of contemporary English*. 2nd Ed. Longman, London.

Sgall P., Hajičová E. and Panevová J. (1986) *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, ed. by J. L. Mey, Dordrecht, Reidel – Prague, Academia.