Evaluation of Tectogrammatical Annotation of PDT

First Observations

Eva Hajičová and Petr Pajas

Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
{hajicova, pajas}@ufal.ms.mff.cuni.cz

Abstract. Two phases of an evaluation of annotating a Czech text corpus on an underlying syntactic level are described and the results are compared and analyzed.

1 Introduction: Tectogrammatical annotation of the Prague Dependency Treebank

Tagging of the Prague Dependency Treebank (PDT in the sequel) on the underlying syntactic layer (resulting in tectogrammatical tree structures, TGTSs) is an ambitious task the realization of which must be carefully supervised and regularly evaluated, in order to reach the proclaimed aims and to obtain (theoretically and applicationally) interesting and applicable results. In the present contribution we describe one of the steps in the introduction of large-scale tagging of a very large corpus.

Tectogrammatical tagging of PDT is carried out in two phases:
(i) an automatic preprocessing transforming the analytic tree structures (ATSs) into structures that are half-way to the TGTSs,
(ii) manual "shaping" of the TGTSs into their final forms. The human annotators have at their disposal a manual [1] and there are regular (weekly) instructive sessions.

2 Description of the evaluation experiment

In order to evaluate the quality of the manual and the instructive sessions and to make estimates about the difficulty of the tagging task (as well as to predict the speed of tagging) we have carried out the following experiment:

Three annotators (all linguists with a university-level education, two having a PhD in linguistics) were given one (randomly chosen) sample of (newspaper) text taken from the Czech National Corpus, which consisted of 50 sentences, with their ATSs preprocessed by the automatic preprocessing procedure mentioned above. They had the manual at their disposal and were asked to tag the sentences according to the manual without negotiations among themselves about the unclear issues. The task of the annotators was to check the dependency structure as such and to assign to the particular values of the dependency relations (functors). They also were supposed to check the lexical values of the nodes and to add appropriate lexical values in case they added some node in the TGTS that was deleted in the surface structure and therefore was missing also in the ATS. A special programme was written to compare the results of the three annotators sentence per sentence (actually, word by word (or node by node)) and to summarize some statistical and qualitative results.

After the evaluation of the results and after a thorough discussion during several instructive sessions about the points in which the annotators differed, we have repeated the same task with the same annotators annotating another randomly chosen set of 47 sentences, to compare the results in order to obtain some judgements about the degree of improvements of the quality of tagging and also to make some predictions about the speed.

3 The first round of the experiment

Out of the total of 50 sentences, only for 10 of them all the annotators gave the same TGTSs; since the number of occurrences of dependency relations in these sentences was not greater than 3, this is a negligible portion of the whole set. The distribution of the number of sentences and the number of differences (one difference means that one dependency relation or a part of a label of a node was assigned in a different way by one of the annotators) is displayed in Table 1. Let us note that the number of dependency relations is slightly smaller

Table 1. The distribution of the number of sentences and the number of differences

No. of diff.	1	2	3	4	5	6	7	9	10	11	13	14	20	21	27
No. of sent.	3	4	4	5	3	3	4	1	6	2	1	1	1	1	1

Table 2. Distribution of sentences according to the number of differences in each sentence ignoring the differences in lemmas

No. of diff.	1	2	3	4	5	6	7	8	9	10	11	18	20	21
No. of sent.	6	4	3	7	4	3	3	2	2	2	1	1	1	1

than the number of words in the sentence, due to the fact that more nodes get taken away than newly restored. The total number of occurrences of dependency relations (i.e. edges) in the test set was 720; the number of differences was 290.

Out of this total number of differences, there were 56 differences in lemmas, 64 differences in the determination of which node depends on which node, and 58 differences in the restoration of nodes not appearing in the ATSs but restored - at least by one of the annotator - in the TGTSs. This leads us to the total of 178 differences in other features than the values of the dependency relations, i.e. out of the total of 720 occurrences of dependency relations (edges) in the first set of sentences there were 112 differences in the assignment of the values of these relations, and 64 differences in the establishment of the edges. It is interesting to note that while the trees with difference 0 were more than simple, the trees with the number of differences $N = 1$ were rather complicated (including 13, 9, 18 relations, respectively), and the same holds about $N = 2$ $(12, 5, 9, 15)$, and $N = 3$ $(26, 11, 17, 4)$. The sentences with $N > 10$ were almost all complex sentences with one or more coordinated structures ($N = 11, 13, 20$), with several general participants expressed by a zero morph ($N = 21$), structures with focus-sensitive particles in combination with negation and with general participants ($N = 14$), and a structure with several surface deletions that should be restored in the TGTS ($N = 27$).

The following observations seem to be important:

(a) The dependency relations (i.e. edges) were correctly established in all cases with the following exceptions:
(i) the position of focussing particles.
(ii) the apposition relation was attached differently in one case
(iii) in several cases, the edges for obligatory though (superficially) deletable complementations of verbs were missing with one or two annotators.

Improvements to be done: for (i) and (ii), more specific instructions should be formulated in the manual; (iii) will improve once the basic lexicon is completed with the assignment of verbal frames specifying the kinds of obligatory complementations.

(b) Lexical labels: the differences concerned uncertainties in assigning the value Gen (for a general participant) and 'on' (pronoun 'he' used in pro-drop cases) and cor (used in cases of control). Improvement: these cases are well-definable and should be more clearly formulated in the manual.

(c) Values of dependency relations: The instructions give a possibility to put a question mark if the annotator is not sure or to use alternatives (two functors). The differences mostly concern uncertainties of the annotators when they try to decide in favour of a single value; other differences are rather rare and concern issues that are matters of linguistic discussions.

4 The second round of the experiment

In the second round of the task, we have evaluated the assignment of TGTSs to 47 sentences in another randomly chosen piece of text (again, taken from the newspaper corpus). When analyzing the results, we faced a striking fact (not so prominent in the first round): there was a considerable amount of differences in the shape rather than in the value of the lexical tags, esp. with lemmas of the general participants (Gen vs. gen) and of the added nodes for coreferring elements (Cor vs. cor). Also other differences in the lemmas were rather negligible caused just by certain changes in the instructions for the annotators.

In Table 3, we count all differences and in Table 4 we ignore the differences in lemmas. A comparison of the two Tables e.g. shows that if differences in lemmas are ignored, the number of sentences with the number of differences equal to 0 through 2 increases from 20 to 26, and that the number of sentences with the number of differences greater than 7 decreases from 10 to 5. The total number

Table 3. Distribution of sentences according to the total number of differences in each sentence

No. of diff.	0	1	2	3	4	5	6	7	8	9	10	11	12	14	16	18	29
No. of sent.	5	8	7	4	4	5	2	2	1	1	1	2	1	1	1	1	1

Table 4. Distribution of sentences according to the number of differences in each sentence ignoring the differences in lemmas

No. of diff.	0	1	2	3	4	5	6	7	11	12	14	28
No. of sent.	8	7	11	2	6	4	2	2	2	1	1	1

of occurrences of dependency relations (i.e. edges) in the second test set was 519; the number of differences was 239. Out of this total number of differences, there were 54 differences in lemmas, 1 difference in the assignment of modality, 35 differences in the determination of which node depends on which node, and 43 differences in the restoration of nodes not appearing in the ATSs but restored - at least by one of the annotator - in the TGTSs. This leads us to the total of 133 differences in other features than the values of the dependency relations, i.e. out of the total of 519 occurrences of dependency relations (edges) in the second set of sentences there were 106 differences in the functors.

In contrast to the first round of the experiment, in the second round the trees with differences 0 were comparatively rich in the number of relations they contained; having 2, 2, 4, 9, and 10 relations if all differences are taken into account, and if the differences in lemmas are ignored, the set of sentences without differences is even enriched by sentences with 7, 11, and 17 relations. The trees with the number of differences $N = 1$ were again rather complicated (including 4, 6, 7, 9, 9, 10, 11, 17 relations, if all differences are taken into account), and the same holds about $N = 2$ $(5, 5, 7, 7, 8, 11, 15)$, and $N = 3$ $(8, 8, 11, 13)$. Similarly as in the first round, the sentences with $N > 10$ included differences in the assignment of general participants expressed by zero morphs (this is true about all sentences in this group), and in most of them the same differences were repeated because of the fact that the sentences included coordination.

5 Comparison

To make Tables 1 and 3. comparable, we exclude the number of sentences with $N = 0$: this group was of no importance in the first round because the sentences included there were very poor in the number of relations they contained, but in the second round this figure is rather important because the sentences belonging there are rather complex (having 2, 2, 4, 9, and 10 relations if all differences are taken into account, and if the differences in lemmas are excluded, this set of is even enriched by sentences with 7, 11, and 17 relations). In total, there are 40 sentences taken into account in Table 1 and 42 in Table 3; out of this total number, 19 sentences in the first round contain less than 5 differences, the rest includes more differences; this number improves in the second round, in which there are 28 with less than 5 differences, i.e. an improvement of almost 50%.

(i) There was a considerable improvement in the assignment of the values of the dependency relations if compared with the first round: out of the total of 123 differences, 21 are not real differences because they consist in an assignment of a "double functor" (or a "slashed" value) by some of the annotators and only one of the values of such a double functor by the other(s). The possibility of an assignment of two (or even more) alternatives to a simple node was introduced in order to make it possible for an annotator to express his/her uncertainty in case even the context does not make it clear what particular relation is concerned (eg. ACMP/COND - Accompaniment or Condition; EFF/RESL - Effect of Result; AIM/BEN - Aim or Benefactive). The introduction of the slashed values is very important for the future research in the taxonomy of dependency relations (merging two current types into one, or making more distinctions) based on the corpus, or formulating the criteria for the distinction between particular values in a more explicit way. In any case, however, the agreement between the annotators on one of the values (and the disregard of the other value by other annotators) should not be really counted as a difference.

There remain, of course, differences which have to be reduced in the further course of the annotation task. The following observations seem to be important for the future development:

(i) As already noticed in (a)(iii) in Sect. 3 above, in several cases, the edges for obligatory though (superficially) deletable complementations of verbs or nouns were missing with one or two annotators. There has been a considerable improvement over the first round since the instructions in the manual have been made more precise in that the restoration of deletable complementations of nouns is restricted to deverbatives in the strict sense, specified by productive derivational means (endings such as *-ání, -ení*). However, there still were cases where the annotators added nodes in cases which were excluded by the instructions (*prodejce* 'seller', *rozhodnutí* 'decision') or were not certain if they are supposed to distinguish two meanings of the deverbative (*uznání* 'recognition' or 'recognizing', *plánování* 'planning' or 'the result of planning'). This is really a difficult point and we may only hope that a better routine will be acquired by the annotators during the annotation process.

(ii) Another case of incorrect restoration is connected with the different types of 'reflexive' forms in Czech. In the TGTSs, a distinction should be made between cases where the reflexive form of the verb is equivalent to a passive (the so-called reflexive passive is very frequent especially in technical texts) , or whether the particle '*se*' is an integral part of the verb (esp. the so-called reflexivum tantum). Examples of the former type occurring in our sample are the forms *šlo se* 'they went', *vytváří se* '(it) is created'; in these case, the lemma '*se*' of the corresponding node in the ATS is 'rewritten' to the lemma of a general participant (gen) and gets the functor of Act; the subject of the (surface) construction gets the functor Pat. In the latter type of reflexive verbs, the ATS node with the label '*se*' is deleted and the particle '*se*' is added to the lexical label of the verb; this is the case of verbs such as *specializovat se* 'specialize', *orientovat se* 'orientate oneself', *představit* si 'imagine', *pustit se* 'to get involved in', *zabývat se* 'occupy oneself with'. Improvement should be reached by more explicit (and more thoroughly exemplified) instructions in the manual.

(iii) Another considerable improvement concerned the cases of lexical labels assigning the value Gen (for a general participant), '*on*' (pronoun 'he' used in pro-drop cases) and cor (used in cases of control). The trivial mistake in the outer shape of the labels (Gen or gen, Cor or cor) will be removed by a macro assigning these values automatically.

(iv) A similar unifying measure should be taken for cases of the assignment of lemmas for pronouns (the lemma of a personal pronoun should be assigned also in cases of a possessive use of pronouns), for the assignment of lemmas to nodes representing certain (meaningful) punctuation marks, and for adding 'empty verbs' in cases when this is necessary for an adequate acount of the dependency structure of the sentence.

To conclude, we believe that the results of both rounds of the evaluation and their comparison will help us to improve the manual for the further process of annotation in order to give better specifications and to help to speed up the work of the annotators. We have also gained several stimuli for linguistic research in areas that have not yet been adequately described in any Czech grammar.

Acknowledgements. We would like to thank our colleagues Jan Hajič for giving us an impetus for the evaluation experiments, Alena Böhmová for providing technical help in carrying them out, and Petr Sgall for an intensive and continuous collaboration on writing and modifying the manual. - Research for this paper was supported mainly by the grant of the Czech Ministry of Education VS96151 and partly also by the grant of the Czech Grant Agency GAČR 405/96/K214.

References

1. Hajičová E., Panevová J., Sgall P.: Manuál pro tektogramatické značkování, ÚFAL Technical Report TR-1999-07, Universitas Carolina Pgragensis (1999)