Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations

Eva Hajičová, Petr Pajas and Kateřina Veselá

Abstract

We summarize here the results of a series of evaluations of the annotators' assignments of tectogrammatical (i.e. underlying syntactic) tree structures and of the values of the edges as well as the values of the attribute representing the topic-focus articulation of the sentences, within the large-scale project of the Prague Dependency Treebank.

1 Introduction: Tectogrammatical Annotations in the Prague Dependency Treebank

Annotations in 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, including a high degree of the automation of the annotating procedure. In the present paper, we describe, compare and analyze results of two phases of an evaluation of annotating a Czech text corpus on this underlying syntactic level (Sections 2 and 3) and we present some preliminary observations on annotators' consistency in the annotation of the basic properties of information structure of the sentences (Section 4).

In the second part of the paper, we present some preliminary observations on annotators' consistency in the annotation of the basic properties of information structure of the sentences (Section 5).

The source texts for the PDT annotation are taken from the Czech National Corpus (CNC), the basic version of which contains 100 million of words from running Czech texts (taken from different language styles). The scenario of PDT is conceived of as a system of three layers, representing the morphemic structure (tagged by a stochastic tagger, see (Hajič and Hladká, 1998)), the surface shape of sentences (resulting in analytic tree structures, ATSs, with 100,000 sentences already annotated, see (Hajič, 1998)), and the underlying (dependency-based) syntactic (tectogrammatical) structure (up to the present point, 20,000 sentences are annotated as for the structure and types of dependency relations and 2,000 sentences are tagged also with the basic features of the information structure (topic-focus articulation, TFA)). A full description of the present stage of PDT, with English translations of manuals and with samples of annotated sentences is available on CD ROM (Hajič et al., 2001).

The tectogrammatical tagging of PDT is carried out in two steps: (i) an automatic preprocessing transforming the analytic tree structures (ATSs) into structures that are half-way to the TGTSs (see (Böhmová, 2001)), (ii) manual 'shaping' of the TGTSs into their final forms. The latter step of tagging proceeds in three 'streams': (a) 'large' corpus is created by a group of six annotators checking the dependency tree structure and the values of the functors (i.e. the labels of the edges of the tree) received from the preprocessing phase, (b) two annotators take the results of the (a) stream and add further values based on a more detailed sub-classification of the dependency relations and some basic values for intersentential relationships; the result

is called the 'model' corpus and for the time being it contains a very small fraction of the large corpus, (c) three annotators add the values of the attribute of topic-focus articulation (see (Hajičová and Sgall, 2001) and Section 4 below) to the sentences from the model corpus.. The annotators have at their disposal a manual ((Hajičová, Panevová, and Sgall, 2000); three 'upgraded' versions have been made available since then) 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 program 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 judgments about the degree of improvements of the quality of tagging and also to make some predictions about the speed.

2.1 The first round of the experiment

Out of the total of 50 sentences, only for 10 sentences all the annotators agreed in the assignment of the TGTSs; since the number of occurrences of dependency relations in these sentences was not higher than 3, the set of 10 sentences with full agreement is a negligible portion of the whole sample. 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 Tables 1 and 2.

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 1: The distribution of the number of sentences and the number of differences

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

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

Let us note that the number of dependency relations is slightly smaller than the number of words in the sentence, due to the fact that during the transition from analytic to tectogrammatical trees the number of nodes taken away is larger than that of the nodes 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:

1. The dependency relations (i.e. edges) were correctly established in all cases with the following exceptions:

   (a) the position of focusing particles

   (b) the apposition relation was attached differently in one case

   (c) in several cases, the edges for obligatory though (superficially) deletable complementations of verbs were missing with one or two annotators.

   These observations have led us to the following measures: for 1a and 1b, more specific instructions should be formulated in the manual; 1c will improve once the basic lexicon is completed with the assignment of verbal frames specifying the kinds of obligatory complementations (a build-up of such a lexicon is described in (Straňáková-Lopatková and Žabokrtský, 2002)).

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

3. 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 favor of a single value; other differences are rather rare and concern issues that are matters of linguistic discussions.

3 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 again the differences in lemmas. A comparison of the two Tables shows e.g. 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.

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 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	11	12	14	28
No. of sent.	8	7	11	2	6	4	2	2	2	1	1	1

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

The total number 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 annotators - 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.

3.1 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 (see above, Sect. 3). 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%.

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 (e.g. 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:

1. As already noticed in 1 1c in Sect. 2.1 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, it still happened that 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.

2. 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* 'one went', *vytváří se* '(it) is created'; in these cases, 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.

3. Another considerable improvement concerned the cases of lexical labels assigning the value Gen (for a general participant), '*on*' (pronoun 'he' 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.

4. 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 account of the dependency structure of the sentence.

Both rounds of the evaluation and their comparison have helped 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.

3.2 The second phase of the evaluation

In order to check the development of consistency in annotation, and also in order to decide at which point a routine annotation phase can be started with the annotators annotating text samples separately rather than in parallel, we followed (after the clarifications in the manual resulting from the first phase of annotation summarized in Sect. 2 had been made) the differences between two (and the same) annotators in the course of annotating six text samples each containing about 50 sentences. The values taken into account were again those in lemmas, structure (i.e. edges of the trees) and functors (i.e. labels of the edges). Since the lemmas again do not seem to bring differences of decisive importance, we merged the differences in lemmas with those of functors on the one hand, and compared them with the differences in the structure. In addition, we relativized the distribution with respect to the proportion of the number of the differences to the total number of compared values: thus e.g. the second column of Table 5 should be read as: in one sentence there was 1% out of all compared values treated differently by the annotators. Tables 5 through 10 exhibit figures for the first of the seven compared samples (each sample contains 50 sentences; the difference of the total number of sentences in the tables showing absolute and relative distribution of differences is given by the fact that some 'sentences' in the sample were empty, i.e. assigned a serial number on technical grounds only).

# diff.	0	1	2	3	4	5	6	7	9
# sent.	14	9	6	8	5	2	3	2	1

Table 5: The absolute distribution of all differences

# diff.	0%	1%	2%	3%	4%	5%	6%8	%	9%	10%	11%	14%	29%
# sent.	11	1	7	3	4	5	6	4	1	1	2	1	1

Table 6: The relative distribution of all differences

# diff.	0	1	2	3	5
# sent.	22	13	13	1	1

Table 7: The absolute distribution of differences in lemmas and functors

A comparison of Tables 3 and 5 shows a significant improvement of consistency (we are aware that the choice of two annotators rather than three contributes to the "plausibility" of the results but hopefully not significantly because our experience indicates that the annotators do not differ in the 'quality' of their assignments): the number of sentences with complete agreement of the annotators has raised from 5 to 14, and the number of sentences with less than 6 differences has raised form 33 to 44. If we accept the assumption that the problems with lemmas have dropped almost to zero and thus compare Tables 4 and 5, we come to similarly agreeable results: the proportion is 8 to 14 of complete agreement, and 38 to 44 for sentences with number of differences less than 6.

# diff.	0%	5%	6%	7%	9%	10%	11%	12%	13%
# sent.	19	4	2	1	3	1	2	1	3

# diff.	14%	15%	17%	28%	20%	22%	25%	50%
# sent.	2	1	1	1	1	1	1	1

Table 8: The relative distribution of differences in lemmas and functors

# diff.	0	1	2
# sent.	45	2	3

Table 9: The absolute distribution of differences in structure

The figures are similar also for the other five samples (see Tables 11 through 15 below): 18, 19, 20, 25, and 15 for zero differences, and 47, 44, 49, 54, and 43 (out of the total number of sentences 50, 53, 55, 54, and 52, respectively). Even more encouraging are the results of the structure assignments: there are 45 (36, 40, 42, 46, 39) sentences in the six samples which have been assigned the same structure by both the annotators. The values of the edges are more difficult to assign: if we assume that the differences in lemmas are negligible and take the figures in Table 7 (and, correspondingly, in the (c) parts of Tables 11 through 15) as an indication of differences in the values of dependency relations, then the number of sentences with a full agreement is 22 (compared to 45 concurrent assignments of structure); similarly for the other samples: 27, 27, 31, 33, 22.

As for the relative distribution, the figures in Table 16 indicate that there is still space for an improvement: the first column in each section displays the percentage of zero difference values of the type concerned out of the total values of that type present in the sample under comparison; the second column gives the percentage for 10% (incl.) difference rate. The rows give the values for the individual samples (sample 1 through 6).

The difference between the absolute and relative distribution confirms one obvious fact that has already emerged from our first evaluation experiment: the annotators do not arrive at an agreement if the structure of the sentence is very complex; the absolute number of agreements is rather high, but in sentences for which the annotators disagree there are more items (be it edges or labels) per sentence with a disagreement. This is first of all the case of coordination, in which the disagreement in one point is multiplied, especially with deletions in the surface shape of the sentence. An improvement of results seems to be beyond the possibility of making the formulations in the manual (and in individual sessions) more precise but lies in our opinion in a rapid and solidly based creation of a huge valency lexicon the development of which is under course (cf. the writings quoted in Sect. 2.1 above) and has already brought some significant improvements in the annotations.

4 Annotating basic features of the information structure (TFA)

4.1 The scheme of TFA annotation

The assignment of basic features of the information structure of sentences (its topic-focus articulation, TFA) is supposed to be an integral part of the TGTSs, which is in accordance to our

# diff.	0%	5%	9%	12%	22%	50%
# sent.	42	1	1	1	1	1

Table 10: The relative distribution of differences in structure

(a) The absolute distribution of all differences

# diff.	0	1	2	3	4	5	7	11	15
# sent.	18	10	13	1	3	2	1	1	1

(b) The relative distribution of all differences

# diff.	0%	1%	2%	3%	5%	6%	7%	8%	12%	18%
# sent.	15	4	9	4	6	1	4	2	1	1

(c) The absolute distribution of differences in lemmas and functors

# diff.	0	1	2	3	4
# sent.	27	15	6	1	1

(d) Sample 3: The relative distribution of differences in lemmas and functors

# diff.	0%	2%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	33%
# sent.	24	1	1	5	1	2	1	1	1	2	5	1	1	1

(e) Sample 3: The absolute distribution of differences in structure

# diff.	0	1	2	3	6
# sent.	36	8	4	1	1

(f) Sample 3: The relative distribution of differences in structure

# diff.	0%	3%	4%	5%	6%	7%	11%	12%	13%	14%	18%
# sent.	33	1	2	1	1	1	4	1	1	1	1

Table 11: Sample 3

empirical observations on the semantic relevance of TFA as well as with the theoretical findings within the framework of the Functional Generative Description (Sgall, Hajičová, and Panevová, 1986). For the moment, the large scale TGTS annotation proceeds in two steps, in the first of which the annotators assign the dependency structure and the labels of the edges (see the sections above) and in the second a different group of annotators adds the values of the TFA attribute and makes changes – if necessary – in the left-to-right placement of the nodes.

At the present stage, we work with a single attribute for TFA offering three values: *t* for a contextually bound (CB) node without a contrast, c for a contrastive CB node, and *f* for a contextually non-bound (NB) node; 'contextual' boundness refers to context in a broader sense of the word, covering both co-text and situation, having also in view that an entity known from the context may be referred to as NB (if presented as new, chosen from a set of alternatives). The CB nodes are placed to the left and the NB nodes to the right of their respective governors, except for the so-called proxy-focus (see below, Sect. 4.2 under b)). The basic instructions for the assignment of the TFA values are specified in the manual and are made more precise in the course of annotation. Up to now 2000 sentences have been annotated in this respect and three annotators are involved in the process: one senior specialist in TFA, one student with a sound linguistic background and one undergraduate. Two senior researchers take part as advisors in the instructive sessions; they also check the results.

4.2 First attempts at an evaluation

In order to check the accuracy of the instructions given in the manual and the consistency of annotators, a preliminary evaluation has been carried out of the annotation of three samples of 50 sentences each by two annotators. The average number of nodes (each node being assigned one of the TFA values) in a sentence is 14; thus the total number of possible differences in the assignments is 2100. The results are given in Table 17.

Discussion: At the first sight, the number of differences is extremely low (37 differences out of approx. 2100 possibilities, i.e. 1,76%); this may be due to the fact that the annotation is carried out by linguists with a reliable knowledge of the TFA framework and to the fact that

(a) The absolute distribution of all differences

# diff.	0	1	2	3	5	6	7	9	16	17	19
# sent.	19	9	7	8	1	3	2	1	1	1	1

(b) The relative distribution of all differences

# diff.	0%	1%	2%	3%	4%	5%	6%	8%	9%	11%	12%	19%	21%	23%	27%
# sent.	16	3	5	6	2	6	3	2	1	1	1	1	1	1	1

(c) The absolute distribution of differences in lemmas and functors

# diff.	0	1	2	3	5	6	7	8	9
# sent.	27	14	4	2	1	2	1	1	1

(d) The relative distribution of differences in lemmas and functors

# diff.	0%	4%	5%	6%	7%	8%	9%	10%	12%	13%	19%
# sent.	24	1	1	2	4	3	1	1	2	2	1

# diff.	22%	23%	28%	29%	33%	38%	55%	85%
# sent.	1	1	1	1	1	1	1	1

(e) The absolute distribution of differences in structure

# diff.	0	1	2	3	6
# sent.	40	9	1	1	2

(f) The relative distribution of differences in structure

# diff.	0%	4%	6%	7%	8%	9%	11%	20%	35%	50%
# sent.	37	1	2	2	3	1	1	1	1	1

Table 12: Sample 4

the repertoire of TFA values is very poor (3) if compared with the repertoire of functors (40). At the same time, a closer examination of the samples reveals that it is rare to find more than one difference in one sentence, so that the total number of differences more or less equals the number of sentences with TFA assigned differently for one of its elements. The intricacies of the sources of differences indicated in Table 17 may be illustrated by the following examples:

a) Contrastive topic: Contrastive topic is basically understood as a choice from a set of alternatives in the topic part of the sentence. The specification of such a set is often left without an explicit specification as in ex. (1).

(1) (*Dva ze skinheadů byli odsouzeni k ročnímu podmíněnému trestu odnětí svobody. Ostatní útočníci byli osvobozeni.*)

(Two of the skinheads were sentenced to a one-year punishment in jail. Other attackers were set free.)

Lit.: Counsellors.t/c at law-suit argued by-the-fact that appurtanence to movement of-skinheads cannot-be determined by outer signs.

Obhájci.t/c při procesu argumentovali tím, že příslušnost k hnutí skinheads nelze určit podle vnějších znaků.

Counsellors at the law-suit argued that the appurtanence to a skinhead movement cannot be determined by outer signs.

In ex. (2) the annotators hesitated between determining the given element as a contrastive CB or an NB node: 'stock-exchange' is a specifier of the noun 'price' and occurs in the text for the first time so that it should be considered to be NB and assigned f; however, it contrasts with the node RM-S and has the same position in the sentence as the latter node (which is duly assigned c); this might indicate that 'stock-exchange' stands in contrast (the first judgment is more appropriate).

(a) The absolute distribution of all differences

# diff.	0	1	2	3	4	5	6	7	9	14	21
# sent.	20	14	8	2	1	4	1	2	1	1	1

(b) The relative distribution of all differences

# diff.	0%	1%	2%	3%	4%	5%	6%	7%	9%	10%	11%	12%	15%	29%	35%
# sent.	6	3	8	3	5	4	2	2	1	1	1	1	1	2	1

(c) The absolute distribution of differences in lemmas and functors

# diff.	0	1	2	3	4	5	9
# sent.	31	11	8	2	1	1	1

(d) The distribution of relative differences in lemmas and functors

# diff.	0%	4%	7%	8%	9%	10%	11%
# sent.	27	1	1	2	1	3	2

# diff.	12%	14%	15%	16%	20%	25%	26%	42%	50%
# sent.	1	2	1	1	3	3	1	1	1

(e) The absolute distribution of differences in structure

# diff.	0	1	2	3	4	11
# sent.	42	7	2	1	2	1

(f) The relative distribution of differences in structure

# diff.	0%	4%	8%	10%	11%	12%	13%	20%	21%	25%	28%	50%	52%
# sent.	38	1	1	2	1	1	1	1	1	1	1	1	1

Table 13: Sample 5

(2) *Zaváděcí cena jeho akcií na pražské burze.f/c byla stanovena na 3300 korun a v RM-S na 1199 korun.*

Lit.: Introductory price of-his shares on Prague stock-exchange.f/c was fixed to 3300 crowns and in RM/S to 1199 crowns.

The initial price of his shares on the Prague stock-exchange was fixed to 3300 crowns and in RM/S to 1199 crowns.

The third example belonging to this class of differences concerns cases with embedded (dependent) clauses in the front position; in Czech, the front position prototypically indicates that the head node of the clause is CB, or contrasted CB. However, some adverbial clauses (esp. those expressing cause or regard) bring a contextually non-bound information and come close to coordination rather than subordination in the semantic interpretation. Then it is not clear whether a choice of alternatives in focus (appropriate with coordination) or in topic (with subordination) is concerned.

(3) *Jestliže se při obchodování akciemi chceme dostat.f/c na standardní způsob jejich převodů, bude nezbytné změnit centrální způsob evidence?*

Lit.: If Refl. at trading with-shares we-want to-get on standard way of-their transfer, it-will-be necessary to-change central way of-registration...

If we want to achieve a standard way of the transfer of shares when trading them, it will be necessary to change the centralized way of the registration...

b) Proxy focus: The notion of proxy focus is introduced for cases of non-prototypical (marked) positions of CB nodes (for more details, see (Hajičová, Partee, and Sgall, 1998)). If a dependent node referring to an entity mentioned in the previous context occurs in the focus part of the sentence and is a governor of focus proper (i.e. the element carrying the intonation center in the surface shape of the sentence), it is difficult to decide whether this node is CB

(a) The absolute distribution of all differences

# diff.	0	1	2	3	4	5
# sent.	25	9	11	3	4	2

(b) The relative distribution of all differences

# diff.	0%	1%	2%	3%	4%	5%	6%	7%	8%	11%	12%
# sent.	20	5	4	3	4	4	5	1	1	1	1

(c) The absolute distribution of differences in lemmas and functors

# diff.	0	1	2	3
# sent.	33	12	8	1

(d) The relative distribution of differences in lemmas and functors

# diff.	0%	5%	6%	9%	10%	11%	12%	14%	16%	18%	20%	25%
# sent.	28	2	5	1	1	3	2	1	1	1	1	3

(e) The absolute distribution of differences in structure

# diff.	0	1
# sent.	46	8

(f) The relative distribution of differences in structure

# diff.	0%	5%	6%	7%	9%	50%
# sent.	41	1	4	1	1	1

Table 14: Sample 6

(and thus constitutes a proxy focus) or (according to its position in the structure) NB. This is illustrated in ex. (4) by the difference in the TFA assignments to the node 'car': it is a governor of the coordinated structure for 'road and terrain' (focus proper) and it is mentioned several times in the preceding co-text; however, here it refers to "a car in general", while in the preceding context a specific type of car is meant; therefore, the assignment of f seems to be more appropriate.

(4) *(Při nedávném představení tohoto vozu novinářům jsme měli možnost získat první dojmy o tomto . . . zajímavém voze.) Chtěli jsme vůz.t/f pro silnici i terén. . .*

(At the occasion of the introduction of this car to the journalists we had the opportunity to get first impressions about this . . . interesting car. . .)

Lit.: We wanted car.t/f for road and-also terrain. We wanted a car well suited both for road and terrain.

Other frequent cases of proxy-focus are constructions involving indications of units of different kinds, as *koruna* 'crown' in ex. (5). The names of units are governors of nodes presenting a substantially more important information, namely the volume (price, value etc.) and they can often be deduced from the co-text (e.g. the price of shares on Prague stock exchange is supposed to be quoted in crowns). However, if the kind of units is mentioned in the sentence for the first time, there seem to be no reasons to classify it as CB. Thus in (5), an assignment of f might be more appropriate.

c) Co-text found in headlines: Since not only the co-text, but also the broader (situational) context is relevant for the determination of CB/NB character of the sentence elements, it is difficult to formulate explicit objective instructions for this determination. In case of article headlines, for instance, a general guiding principle is that they are considered as a part of the co-text; however, the headline may also be just a summary of the contents of the article and as such lies 'outside' the text that follows.

In ex. (6), the difference concerns the assignment of the value t in the former interpretation, and of the value f when the annotator considered the headline to be a summary.

(a) Sample 7: The absolute distribution of all differences

# diff.	0	1	2	3	4	5	6	7	8	9	10	11
# sent.	15	13	6	4	2	3	2	2	2	1	1	1

(b) Sample 7: The relative distribution of all differences

# diff.	0%	1%	2%	3%	4%	5%	6%	7%	8%	9%	11%	12%
# sent.	12	6	7	3	4	2	4	3	2	3	2	1

(c) Sample 7: The absolute distribution of differences in lemmas and functors

# diff.	0	1	2	3	4
# sent.	22	19	4	6	1

(d) Sample 7: The relative distribution of differences in lemmas and functors

# diff.	0%	3%	4%	5%	6%	7%	8%	9%
# sent.	19	3	2	2	1	1	3	3

# diff.	10%	11%	12%	14%	15%	16%	25%	50%
# sent.	2	2	2	3	1	2	2	1

(e) Sample 7: The absolute distribution of differences in structure

# diff.	0	1	2	3	4	5
# sent.	39	7	1	1	3	1

(f) Sample 7: The relative distribution of differences in structure

# diff.	0%	4%	5%	6%	7%	10%	12%	14%	15%	17%	18%
# sent.	36	2	2	1	1	1	1	1	2	1	1

Table 15: Sample 7

Values of functors (labels)		**Edges (structure)**	
0%	**below 10% incl.**	**0%**	**below 10% incl.**
19	30	42	44
24	37	33	39
24	37	37	46
27	35	38	42
28	37	41	48
19	36	36	43

Table 16: The relative distribution of discrepancies in individual samples

(6) *(Hedaline: Žižkov zopakoval nevydařený start z jara.) Překvapením kola je prohra Viktorie Žižkov.t/f v Hradci Králové.*

(Headline: Žižkov repeated its unsuccessful start from the spring.)

Lit.: Surprise of-round is defeat of-Victoria Žižkov in Hradec Králové. The surprise of this round is the defeat of V.Ž. in H.K.

d) Other cases: This group comprises mostly hesitations of the annotators in considering the given node as CB (and thus to be assigned the values t or c) or NB (value f). Though the instructions in the manual offer a rather detailed guide, there is always some space for the annotators' individual understanding and judgments. Also, it is quite natural that in some cases the annotators failed to take the details of the instructions into account.

4.3 Tentative conclusions

The TFA annotation is still in its first, experimental stage and a comparison of the annotators' assignments, though still rudimentary, is extremely instructive and helpful in many respects: the

Total no. of sentences	150
Total no. of differences	37
Differences due to different understanding of:	
(a) contrastive topic	11
(b) proxy focus	8
(c) co-text in headlines	10
(d) other reasons	8

Table 17: Results of preliminary evaluation of TFA annotation

manual should be carefully complemented by more detailed and clear instructions and examples; the issues that have appeared to be fuzzy have to be studied more closely both empirically and theoretically (this is especially the case of the notions of contrastive topic and of proxy-focus), and the annotation should still continue in parallel, with a comparison and evaluation of the results. At the same time, it seems to be a topical (and feasible) issue already now to try and implement an algorithm for an automatic preprocessing of the TGTSs with structural annotations (reached in the first 'stream', see Sect. 1 above) in order to assign automatically the TFA values to the core of them (for a first tentative formulation of such an algorithm, see (Hajičová, Skoumalová, and Sgall, 1995)). One point, however, is clear already now: the results of the evaluation, even if carried out on a small sample, are encouraging and confirm that annotation of information structure is feasible and brings important stimuli for further linguistic and computational research.

5 Summary

Our experience has confirmed the usefulness of the evaluation of the annotations as for the consistency of the annotators (which, of course, depends significantly on the consistency of the instructions they are given). The evaluation has helped us to make the instructions more precise, which included also a more detailed study of several language phenomena that have not yet been sufficiently treated in the existing grammars of Czech, as well as to speed up the annotation by developing some additional annotating tools or macros and by allowing the annotators to annotate different samples each, with one annotator going through all the annotated samples and checking them, again with the help of specific software tools.

Acknowledgments

We would like to thank our colleagues Jan Hajič for giving us the impetus for the evaluation experiments, Alena Böhmová for providing technical help in carrying them out, and Petr Sgall for most useful pieces of advice and comments as well as for an intensive and continuous collaboration on writing and modifying the manual.

Research for this paper was supported mostly by the grant of the Czech Ministry of Education LN00A063 and partly also by the grant of the Czech Grant Agency GACR 405/96/K214.

References

Böhmová, Alena. 2001. Automatic procedures in tectogrammatical tagging. *Prague Bulletin of Mathematical Linguistics*, 76.

Hajič, Jan. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*. Prague Karolinum, Charles University Press, pages 12–19.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. 2001. Prague dependency treebank 1.0 (final production label). CDROM CAT: LDC2001T10., ISBN 1-58563-212-0.

Hajič, Jan and Barbora Hladká. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pages 483–490, Montreal, Canada.

Hajičová, Eva, Jarmila Panevová, and Petr Sgall. 2000. A manual for tectogrammatic tagging of the prague dependency treebank. Technical Report TR-2000-09, ÚFAL MFF UK, Prague, Czech Republic. in Czech.

Hajičová, Eva, Barbara Partee, and Petr Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer Academic Publishers, Amsterdam, Netherlands.

Hajičová, Eva and Petr Sgall. 2001. Annotating the basic features of the information structure (tfa). In *Proceedings of ACL 2001*, Toulouse, France.

Hajičová, Eva, Hana Skoumalová, and Petr Sgall. 1995. An automatic procedure identifying topic and focus. *Computational Lingusitics*.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Straňáková-Lopatková, Markéta and Zdeněk Žabokrtský. 2002. Valency dictionary of czech verbs: Complex tectogrammatical annotation. In *LREC 2002 Proceedings*, Las Palmas, Gran Canaria.