Annotators' Agreement: The Case of Topic-Focus Articulation

Kateřina Veselá, Jiří Havelka and Eva Hajičová

Center for Computational Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 11800 Prague 1, Czech Republic
{vesela, havelka, hajicova}@ckl.mff.cuni.cz

Abstract

The annotation of the Prague Dependency Treebank (PDT) is conceived of as a multilayered scenario that comprises also dependency representations (tectogrammatical tree structures, TGTS's) of the underlying structure of the sentences. TGTS's capture three basic aspects of the underlying structure of sentences: (a) the dependency tree structure, (b) the kinds of dependency syntactic relations, and (c) the basic characteristics of the topic-focus articulation (TFA). Since the PDT is a large collection and the annotations on the deepest layer are to a large extent performed by several human annotators (based on an automatic preprocessing module), it is more than necessary to observe the consistence of annotators and the agreement among them. In the present paper, we summarize the results of the evaluation of parallel annotations of several samples taken from PDT and the measures accepted to improve the consistency of annotations.

1. Introduction

The deep syntactic structure annotations in the Prague Dependency Treebank (PDT), the so-called tectogrammatical tree structures (TGTS's), capture, among other things (such as a preliminary assignment of coreference), three basic aspects of the underlying structure of sentences: (a) the dependency tree structure, (b) the kinds of dependency syntactic relations, and (c) the basic characteristics of the topic-focus articulation (TFA). Since the PDT is a large collection (the ultimate aim is to annotate 100,000 Czech sentences on three levels of depth, namely the morphemic layer, the surface shape of sentences and tectogrammatical layer), and the annotations on the deepest layer are to a large extent performed by several human annotators (based on an automatic preprocessing module), it is more than necessary to observe the consistence of annotators and the agreement among them. In the present paper, we summarize the results of the evaluation of parallel annotations of several samples taken from PDT and the measures accepted to improve the consistency of annotations.

2. Annotation of Topic-Focus Articulation

2.1 The Theoretical Basis of the TFA Annotation

The build-up of the tectogrammatical tree structures in PDT is based on the formal framework of Functional Generative Description (FGD; for its background and basic notions see Sgall et al., 1986); the theoretical framework offers also a very consistent and formally sound account of topic-focus articulation of sentences (TFA; information structure).

In FGD, the semantic basis of the articulation of the sentence into T(opic) and F(ocus) is the relation of aboutness: a prototypical declarative sentence asserts that its F holds (or does not hold) about its T: F(T) or non-F(T). Within both T and F, an opposition of contextually bound and non-bound nodes is distinguished, which is understood as a grammatically patterned opposition, rather than in the literal sense of the term. Within the contextually bound elements of the sentence, a difference is made between contrastive and non-contrastive bound elements. Hajičová et al. (1998, p. 151) introduce the notion of contrastive (part of) T in connection with the occurrences of the so-called focusing particles in T (such particles as *only, even, also* etc.); they use the index c to mark the item in such a position; however, in the course of our further empirical investigations we have found a clear evidence that contrast in T is not connected only with the occurrences of focusing particles.

Example (1), taken from PDT, illustrates the typical layer (the sentence is supposed to be pronounced with an unmarked position of the intonation center, i.e. with its placement at the end of the sentence).

Notational convention for the example: Since the function words such as prepositions and auxiliary verbs do not have a node of their own on the underlying level of FGD, they are in our schematic notation (i.e. in the primed example) included in brackets. The index b denotes the given element as contextually bound, elements with no index are considered to be contextually non-bound; the index c denotes the given element as a contrastive contextually bound element. The elements of F of the Czech sentences are denoted by italics.

(1) (V) noci (ze) soboty (na) neděli skončil (ve) vojenském prostoru Ralsko sjezd majorů.
Lit. E. transl: (At) night (from) Saturday (to) Sunday ended (in) military area Ralsko meeting-Nom. of-majors.
Question: What happened during the night from Saturday to Sunday?
T: v noci ze soboty na neděli
F: skončil ve vojenském prostoru Ralsko sjezd majorů
(1') (V) noci.b (ze) soboty.b (na) neděli.b *skončil (ve) vojenském prostoru Ralsko sjezd majorů.*

2.2. Corpus Annotation with respect to TFA

The tectogrammatical tree structures (TGTS's) capture the syntactic (dependency) relations, such as ACTor, ADDRessee, Objective (PATient), LOCative, DIRection, MANner, restrictive attribute (RSTR), RHEMatizer, etc., and morphological values, such as Preterite (Anterior), Conditional, Plural, etc., and also the prototypical values of 'in', 'into', 'on', 'from', etc. They describe also the

TFA of the utterances in the corpus, since TFA is expressed by grammatical means and is relevant for the meaning of the sentence (even for its truth conditions), i.e. it constitutes one of the basic aspects of underlying structures (for arguments on the semantic relevance of TFA, see e.g. Sgall et al., 1986; for the relevance of TFA for the semantics of negation, see Hajičová, 1984). The nodes in the tree are ordered according to the degrees of communicative dynamism (deep word order).

The following three values of the attribute TFA are distinguished with every node in a TGTS:

(i) T: a non-contrastive CB node, which always has a lower degree of CD than its governor;
(ii) F: an NB node (if different from the main verb, then following after its head word in the TGTS);
(iii) C: a contrastive CB node.

Example (2) and the corresponding (rather sketchy) TGTS in Figure 1 illustrate the result of the TFA assignments:

(2) Už první pohled na atypickou karosérii potvrzuje, že se jim podařilo tento záměr naplnit.
Lit. E. transl.: Already first look at atypical car-body confirms, that Refl. them succeeded this intention to-fulfil.
E. transl.: Already the first look at the atypical car-body confirms that they have succeeded in meeting the intention.

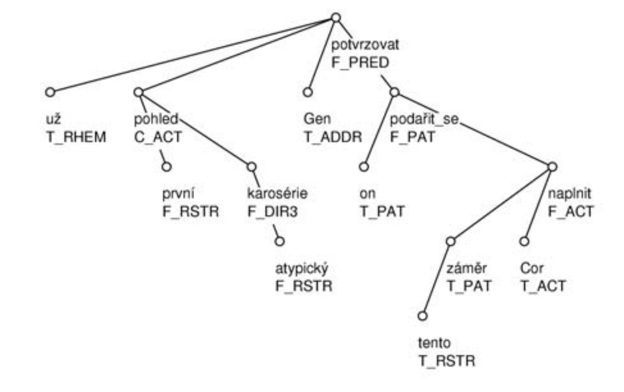

Figure 1: A sample TGTS

Gen and Cor are formal lemmas of nodes restored in the TGTS's (i.e. their correlates are absent in the surface shape of sentences).

The division of the sentence into topic and focus can be derived on the basis of the assignments of the TFA values and it corresponds to the context in which the sentence occurs in the annotated text.

3. Evaluation of the Annotation of TFA

3.1. Course of Annotations

TFA is being annotated within the PDT project since 2001. The sentences that are to be annotated on TFA have been syntactically analyzed and the types of syntactic dependencies have been marked – the annotator of TFA assigns values T, F or C to nodes and modifies the deep order in the dependency tree. In 2001 and 2002 a so-called test annotation was carried out and since 2003 the annotators of TFA have been annotating data in their definitive form (the PDT).

The annotation is based on the Manual for Tectogrammatical Tagging of the Prague Dependency Treebank. In the course of annotation the annotators held meetings for solving problems which they had encountered in the texts but which either were not covered by the Manual at all, or were not elaborated in enough detail. From 2002 parallel annotations have been performed – approximately every six months (spring 2002, autumn 2002, spring 2003, autumn 2003) a sample of data was chosen and was annotated in parallel. The results of a theoretical analysis of the arising disagreements served as a basis for delimiting problematic issues, a draft of their solution in the annotation, and eventually the elaboration of the corresponding section of the manual. By the end of 2003 the development of the manual for the annotation of TFA was closed (Veselá and Havelka, 2003); following this, a comparison of several files was carried out in order to find out whether there remained any phenomena whose inconsistent annotation could lead to a considerable amount of errors in the data. From the last two years we have therefore at our disposal data which allow us to evaluate to a certain extent the evolvement of the annotation from the point of view of the agreement between annotators, as well as the phenomena resulting in annotation disagreements. Our main goal is to describe and classify the cases in which the annotators disagree, in order to be able to concentrate our work on the annotation of TFA in this direction. The overall amount of agreement is also a good indicator of the relevance of newly introduced guidelines.

3.2. Input Data

There were four annotators involved in the annotation of the data used for the evaluation – an experienced annotator (annotator 1) and three students of the Czech language (annotators 2, 3 and 4). The annotation was started by annotators 1, 2 and 3, at the end of 2002 annotator 1 left and was replaced by annotator 4. Therefore we can compare three versions of annotation for each file concerned. The whole data set consists of 441 triples of annotated sentence structures. Table 1 gives an overview of the files annotated in parallel and their respective sizes in the number of trees and nodes.

Phase	# files	# trees	# nodes	Annotators
1 (spring 2002)	2	94	1338	1, 2, 3
2 (autumn 2002)	1	48	825	1, 2, 3
3 (spring 2003)	1	52	702	2, 3, 4
4 (autumn 2003)	5	247	3537	2, 3, 4
Total	9	441	6402	

Table 1: Data annotated in parallel

All the numbers have been obtained using specific computational tools and subsequently manually checked and classified. The classification criteria and procedures will be described in corresponding subsections.

4. Results and Discussion

4.1. Overall Results

Table 2 presents overall results for the agreement between annotators. The agreement in the annotation on nodes for the first two phases is about 80%. After the substitution of one annotator the degree of agreement somewhat decreases, but it sharply increases in the fourth phase – up to 90%.

Phase	1	2	3	4
Agr. on nodes	81.32	81.89	76.21	89.57
Agr. on all trees	25.53	29.17	28.85	35.63
Agr. on relevant trees	20.00	20.93	24.50	29.02

Table 2: Agreement between annotators (percents)

The second and the third lines of Table 2 give the percentages of agreements if trees as a whole are taken into account; the percentage of agreements drops down if only sentences exhibiting a true topic-focus structure are counted (PDT texts are very diverse and include many types of non-sentential constructions). However, a continual improvement can be observed even there.

The discrepancy between an 80–90% agreement on nodes and a 30% agreement on trees means that although most of the nodes can be annotated quite unambiguously, the phenomena causing disagreement cannot be denounced as marginal, because they play a role in almost three fourths of all annotated trees. In the next subsection we attempt to delimit and further analyze these issues.

4.2. Problem Areas

Table 3 gives a list of particular theoretical areas that we have especially focused on. We have taken into account only those cases where the annotators disagreed between two values. We consider the cases where each annotator used a different value for contextual boundness to be too unclear, moreover they occur very scarcely in the data (they constitute in each phase less than one percent of all cases). The percentages in the table say how frequent individual phenomena are relative to all disagreements.

Phase	1	2	3	4	Total
1. Nodes with C	9.20	2.67	8.38	5.96	6.73
2. Position of C	0.80	0.67	1.80	3.25	1.92
3. Adjuncts with C	4.80	2.00	3.59	4.88	4.17
4. Unit expressions	4.00	6.67	4.79	3.52	4.38
5. Attributes	36.40	48.67	27.54	34.15	35.90
6. Settings	8.40	10.67	8.98	11.38	10.04
7. Errors	3.60	7.33	8.98	2.98	4.91
8. Other cases	32.80	21.33	35.93	33.88	31.94
9. C/T v. F	72.40	82.00	62.87	69.65	71.15
10. C v. T	22.00	11.33	33.53	28.18	24.79

Table 3: Problem areas and proportions of disagreements

Individual disagreements in annotation are not considered as errors, but as potential values of contextual boundness – we never can accurately determine the contextual boundness of a node, because we cannot dispose of ample enough (complete) context for a univocal decision. Therefore, the values of contextual boundness of individual nodes are only to be taken as more or less probable. The distribution of diverging annotations of a node serve to evaluate these tendencies.

4.2.1. Introductory Remarks

Most valuable for us are the observations reflected in lines 9 and 10 of Table 3. Line 9 summarizes the proportion of disagreements in the assignment of T and C on one hand, and F on the other; the proportion of disagreements between C and T relative to all disagreements can be found in line 10.

Contrastive topic has been introduced into the theory only recently, and we are not yet able to determine all its properties and distributional characteristics. At the beginning its tagging was based to a certain extent on intuition, more detailed instructions were being developed only using the problematic issues encountered during the course of annotation. This has led us to pay an even increased attention to contrastive topic in our evaluation.

It is important to note that the disagreements between C and F bear on a different problem – since focus can be understood as always involving some kind of contrast, a contextually bound item carrying a contrastive feature can be easily misunderstood as a part of focus. In spoken language, prosody can be taken as a helpful criterion (see Veselá et al., 2003).

4.2.2. Contrastive Topic

The notion of contrastive topic as a category is based upon the semantic relation of contrast of individual nodes to nodes in the preceding context. However, there are several other factors at play. To set apart cases where the decision taken is based only on the contrastive relation of a node to its context, we selected all nodes depending directly on a verb and not governing any other node – in such cases we can be fairly sure that the problems in annotation are not caused by confusion about the syntactic structure. As we can see in line 1 of Table 3, compared to line 10, these cases form about one fourth of all cases of disagreement in the annotation of contrast.

Apart from these simple cases we counted also cases with a more complicated structure. We compared dependency edges where the annotation of contrast is shifted – in one annotation the governing node of the edge is marked as contrastive, in another one the depending node is marked as such. The resulting numbers (line 2) might seem low, but the occurrence of such edges in the annotated text is quite rare. In the fourth phase, in which a larger amount of text was annotated, the number of cases where annotators hesitated about the position of contrast in the dependency structure is nevertheless not negligible.

As far as the evolvement of annotation of contrastive topic (see line 10) is concerned, we can see that the most important discrepancies were in the first and third phases of annotation, and that they decreased substantially in both teams of annotators. Contrastive topic however still remains a fundamental “debt” of our annotation guidelines.

4.2.3. Syntactic Functions of Nodes

The syntactic functions of nodes also play a role in the annotation of TFA. We concentrated on free modifiers or adjuncts (of place and time); see lines 3 and 6 of Table 3. Due to the tendency in Czech to place the verb in the second position in the sentence, it is quite usual that contextually bound adjuncts of place and time (so-called settings) occur after the verb and it is thus difficult to decide about their contextual boundness, even more so in the case of concrete specifications. Therefore we wanted to check the proportion of disagreements in free modifiers relative to all disagreements: according to line 6 they form about one tenth of all disagreements between T and F.

We also examined for free modifiers the disagreements between C and T – these should not be too important, because it is arguments (inner participants) that tend to be contrastive (due to their semantic relevance). The results nonetheless are not as low as we expected and they again show the difficulties in the definition of the notion of contrast.

4.2.4. Nominal group

Lines 4 and 5 of Table 3 concern agreement in annotation within nominal groups. As the word-order in nominal group in Czech is more or less fixed and the position of the local intonation center is often unclear, it is hard to establish the communicative dynamism within a nominal group. The only clues can be found in verbatim repetitions of items from the context, also the grammatical realization of attributes can be helpful. In the manual for the annotation of TFA the instructions for the annotation of nominal groups have been modified several times due to the fact that the degree of disagreement in this issue was high. This was also the reason why we decided to test this kind of disagreement. We concentrated on attributes (we were looking for nodes directly depending on a noun, line 5) and nodes governing numerals (this is a subtype of nominal groups, line 4).

We can see from the results that disagreements in the annotation of contextual boundness of attributes amount to about 40% (in the second phase even half) of the total number of disagreements in the annotation of TFA.

4.2.5. Errors

Although we are not able to determine with absolute certainty the values of contextual boundness, some instructions of the manual are hard and fast – especially in cases of nodes where the values of contextual boundness are assigned arbitrarily, because they are irrelevant or unimportant from the point of view of TFA (e.g. in the case of restored nodes). Such instructions get violated only because of ignorance or distractedness of annotators. Errors (line 7 of Table 3) account for about 5% of disagreements in the data compared, towards the end their amount decreases significantly.

4.2.6. Other Cases

Line 8 of Table 3 tells us that about 30%of all disagreements are not covered by our classification. We can conclude that we were able to determine most of the problems causing disagreement in the annotated data. The remaining cases need further study and classification.

4. 3. Statistics for Individual Annotators

It is clear that disagreements in annotations are heavily influenced by the interpretation of contextual boundness by individual annotators. Table 4 shows in how many cases a particular annotator used a particular value of contextual boundness. The last three lines present the decisions taken by particular annotators in cases of nodes with two different annotations.

Annotator	1	2	3	4
T	722	1652	1754	1012
C	35	379	298	157
F	1140	3602	3576	1976
C (from T/C)	4/72	153/176	79/176	23/104
T (from T/F)	180/291	192/530	227/530	115/239
F (from C/F)	11/13	10/31	15/31	11/18

Table 4: Annotators' statistics

We can observe apparent differences especially in the annotation of contrastive topic between annotators 2 and 3 on one hand, and annotators 1 and 4 on the other. Annotator 1 also uses most often the value T – she tags more nodes as contextually bound.

5. Conclusions

The main problem areas in annotation of TFA in PDT are areas that have not yet been adequately elaborated theoretically – above all communicative dynamism in nominal groups and the notion of contrastive topic.

Overall agreement of approximately 80% seems to be sufficiently high to allow us to conclude that the annotation of TFA is feasible, the perception of contextual boundness is not too subjective to disallow a reliable enough annotation of texts. The substantial increase in agreement towards the end of our evaluation indicates that the completion of the manual for annotation helped to raise the reliability of annotation and that the elaboration of hypotheses and their applications in Functional Generative Description helped the annotators to deeper understand the subject matter and make the annotations of TFA more consistent.

There remain two main tasks left for the future: apart from further study of the above-mentioned theoretical issues also the comparison of the deep ordering of nodes – only this will make our evaluation of the annotation of TFA complete.

Acknowledgements

The research reported in this paper has been carried out under the project LN00A063 supported by the Czech Ministry of Education.

References

Hajičová, Eva, 1984. Presupposition and allegation revisited. *Journal of Pragmatics* 8:155–167.

Hajičová, Eva, Barbara H. Partee, and Petr Sgall, 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer, Dordrecht.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová, 1986 *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

Veselá, Kateřina, Nino Peterek, and Eva Hajičová, 2003. Topic-Focus Articulation in PDT: Prosodic Characteristics of Contrastive Topic. *Prague Bulletin of Mathematical Linguistics*, 79–80:5–22.

Veselá, Kateřina, and Jiří Havelka, 2003. Anotování aktuálního členění věty v Pražském závislostním korpusu (Annotation of TFA in Prague Dependency Treebank). ÚFAL/CKL Technical Report TR-2003-20.