(Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank

Jiří Mírovský, Kateřina Rysová, Magdaléna Rysová, Eva Hajičová
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
{mirovsky|rysova|magdalena.rysova|hajicova}@ufal.mff.cuni.cz

Abstract

The objective of the present contribution is to give a survey of the annotation of information structure in the Czech part of the Prague Czech-English Dependency Treebank. We report on this first step in the process of building a parallel annotation of information structure in this corpus, and elaborate on the automatic pre-annotation procedure for the Czech part. The results of the pre-annotation are evaluated, based on the comparison of the automatic and manual annotation.

1 Introduction

In the past three or four decades, topic-focus articulation (known also as sentence information structure) is a language phenomenon that has attracted an enormous interest in linguistics and has become a "hot" topic of linguistic studies. No wonder then, that these days several linguistic teams (e.g. at the University of Potsdam, University of Berlin, University of Stuttgart, Charles University in Prague) have attempted to include the annotation of information structure in the annotating schemes they propose. Among corpora that contain also annotation of information structure or such type of annotation is planned in them there are e.g. ANNIS database (Annotation of Information Structure, see Dipper et al., 2004), The English Switchboard Corpus (see Calhoun et al., 2005), the corpus DannPASS (Danish Phonetically Annotated Spontaneous Speech, see Paggio, 2006) and the Prague Dependency Treebank (for the information on PDT, see Hajič et al., 2006).

There are also several types of annotation guidelines and schemes for the different corpora, based on various linguistic theories dealing with information structure (e.g. Hajičová et al., 2000; Nissim et al., 2004; Dipper et al., 2007; Donhauser, 2007; Cook and Bildhauer, 2011).

In our paper, we present the annotation of topic-focus articulation in the Czech part of the Prague Czech-English Dependency Treebank, based on the theory of topic-focus articulation as developed withing the Praguian Functional Generative Description. It is the first step in the process of building a parallel Czech-English corpus annotated with this type of linguistic information.[1]

1.1 Topic-Focus Articulation in Prague Treebanks

The first complex and consistent theoretically-based annotation of topic-focus articulation was already fully applied in the first Czech corpus from the Prague corpora family, the Prague Dependency Treebank (PDT; Hajič et al., 2006, updated in Bejček et al., 2012), and is available for the linguistic community. PDT is a large collection of Czech journalistic texts, (basically) manually annotated on several layers of language description (more than 3 thousand documents consisting of almost 50 thousand sentences are annotated on all the levels).

Detailed annotation guidelines that constitute the basis of the handling with the language material were developed (Mikulová et al., 2005) based on the theoretical assumptions of the Functional Generative Grammar (for the first formulations of this formal framework, see Sgall, 1967; Sgall et al., 1986). The annotation of the information structure in PDT is also based on this theory. The same linguistic approach was used in some other annotation schemes connected with the annotation of topic-focus articulation (e.g. Postolache, 2005).

[1] Given the available funds, our present goal is to annotate 5 thousand parallel sentences.

International Joint Conference on Natural Language Processing, pages 55–63,
Nagoya, Japan, 14-18 October 2013.

1.2 Aim of the Paper

Our effort is concentrated on annotating the topic-focus articulation (TFA) in a parallel corpus – the Prague Czech-English Dependency Treebank (PCEDT), to make possible contrastive studies of this phenomenon. As the first step, we annotate topic-focus articulation in the Czech part of the treebank. The annotation guidelines have been taken over from the PDT approach, i.e. they also follow the theory of Functional Generative Description.

In Section 2, we give an overview of the theoretical background of TFA, Section 3 introduces the Prague Czech-English Dependency Treebank (the data to be annotated). Section 4 describes in detail an automatic pre-annotation procedure that was applied on the data before they were annotated manually by a human annotator. The final step of this part of our research was the evaluation of effectiveness of the automatic pre-annotation, given in Section 5.

2 Theoretical Background for Corpus Annotation of Topic-Focus Articulation in PCEDT

The theoretical linguistic background for the creating of the whole corpus PCEDT is the Functional Generative Description (Sgall, 1967; Sgall et al., 1986). Topic-focus articulation in this theoretical framework was described especially by Sgall and Hajičová (summarized in Sgall et al., 1986, Hajičová et al., 1998). On the basis of this, the annotation guidelines for manual annotation of topic-focus articulation in the Prague Dependency Treebank (PDT) were established and are available in the annotation manual for the underlying structure of sentences in Mikulová et al. (2005). These guidelines are used also for the Czech part of the Prague Czech-English Dependency Treebank.

2.1 Topic-Focus Articulation in Functional Generative Description

The theory of topic-focus articulation within the framework of Functional Generative Description is based on the aboutness-principle: the topic is the part of a sentence that is spoken about, and, complementarily, the focus is the sentence part that declares something about the topic. From the cognitive point of view, topic may be characterized as the "given" part of the sentence and focus as the "new" one. However, this does not mean that the focus elements cannot be mentioned in the previous language context at all but they have to bring some non-identifiable information or information in new relations.

Most sentences contain both parts – topic and focus. However, some sentences can be contextually independent (e.g. the first sentence of the text or its title) and they do not have to contain the topic part (these are topic-less sentences). On the contrary, the focus is an obligatory component of every sentence – it is the informatively more important part of the message than the topic.

The basic opposition established by the TFA theory and included in the annotation scheme is the opposition of contextual boundness: each element of the underlying structure of the sentence carries the feature "contextually bound" or "contextually non-bound". In addition, the contextually bound elements in the topic can be either contrastive, or non-contrastive. Contrastive contextually bound sentence members differ from the non-contrastive ones in the presence of a contrastive stress and in their semantic content – they express contrast to some previous context (e.g. *at home – abroad*).

Non-contrastive contextually bound expressions are marked as 't', contrastive contextually bound expressions are marked as 'c' and contextually non-bound expressions are marked as 'f'[2].

The opposition between contextually bound and contextually non-bound elements serves then as a basis for the bi-partition of the sentence into its topic and focus; according to this hypothesis, an algorithm for topic-focus bi-partition was formulated, implemented and tested on the PDT data, with some rather encouraging results (see Hajičová et al., 2005).

In Czech (Czech is the language of Prague Dependency Treebank and also of one half of the Prague Czech-English Dependency Treebank), the word order position of predicative verb is often the natural boundary between the topic and focus part in the sentence – cf. Example (1).

(1) [Context: *Moje matka má ráda růže a tulipány.*] *Tulipány*$_{contrastive\_topic}$ *matka*$_{topic}$ *včera*$_{topic}$ ***koupila***$_{focus}$ ***na trhu***$_{focus}$

Literally: [Context: *My mother likes roses and tulips.*] *The tulips*$_{contrastive\_topic}$ *the mother*$_{topic}$ *yesterday*$_{topic}$ ***bought***$_{focus}$ ***on the market***$_{focus}$.

[2] The contextually non-bound elements do not have a contrastive and non-contrastive variant in the theory of FGP.

(= *The mother bought the tulips ON THE MARKET*[3] *yesterday.*)

Several operational tests have been proposed in literature that help to distinguish between topic and focus, the most relevant of them being the question test and the test of negation (for details see Sgall et al., 1986; Hajičová et al., 1998).

In short, the basis of the question test is to ask a question that fully represents the context for the tested sentence. The tested sentence has to be a relevant answer to the question. The sentence members present in both the question and answer are topic members. The elements present only in the answer are members of the focus.

The principle of the negation test is to find out the possible scope of negation in the negative counterpart to the given sentence. In principle, the sentence members that are in the scope of negation in the given context belong to the focus part of the sentence. Other members form the topic part. However, there is a possibility of negative topic, i.e. the topic of the sentence is negated and the focus stands out of the scope (for details see e.g. Sgall et al., 1973).

For detailed information on annotation guidelines of topic-focus articulation in the framework of Functional Generative Description, the online annotation manual is available (see http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html).

3 Language Material – Prague Czech-English Dependency Treebank

The annotation effort described in this paper is performed on data from the Prague Czech-English Dependency Treebank (PCEDT, Hajič et al., 2012), a manually parsed parallel Czech-English corpus that contains over 1.2 million running words (50 thousand sentences in each of the two languages). The English part consists of texts from the Penn Treebank (Marcus et al., 1993) – articles from the Wall Street Journal. The Czech part contains human translations of the English sentences to Czech.

The annotation (on both language sides) is performed on four language layers: the "word" layer, the morphological layer, the analytical layer (i.e. the layer of surface syntax) and the tectogrammatical layer (i.e. the semantic layer of the deep syntax).

On the topmost (tectogrammatical) layer, individual sentences are organized in dependency tree structures, according to the style of the Prague Dependency Treebank (PDT). Autosemantic words and coordinating structures are captured in the trees, as well as the valency of verbs (each language has its own valency lexicon in PCEDT). Additionally, the surface sentence ellipsis is reconstructed in the deep sentence structure and also pronominal anaphoric relations are labeled in the texts. The topic-focus articulation is also to be annotated on this layer.

The parallel Czech-English data are aligned manually on the level of sentences and automatically on the level of tectogrammatical nodes.

More detailed information on PCEDT is available on the project website (http://ufal.mff.cuni.cz/pcedt2.0/en/index.html).

4 Automatic Pre-Annotation

For the annotation of topic-focus articulation in the Czech part of PCEDT, an automatic pre-annotation procedure was developed. The particular steps (rules) of the pre-annotation were mainly established on the basis of the completed annotation of contextual boundness in the Prague Dependency Treebank (i.e. on the basis of annotated Czech texts). The cross-language alignment of tectogrammatical nodes in PCEDT was also exploited (see the pre-annotation step 10 below), allowing for taking advantage of the existence of indefinite articles in English (not present in the Czech language).

Using information from the English side for the pre-annotation of topic-focus articulation in the Czech part is possible, as the topic-focus articulation of the given sentence in the given context should be identical regardless on the language[4]. The surface word order may vary in Czech in comparison with English (cf. the different word order in Example (1) in the two languages) but the topic-focus articulation of the sentence should be the same in both the languages. This theoretical assumption, as well as the quality of the English->Czech translation (from the point of view of topic-focus articulation), can be tested on real corpus data once the annotation on both language sides of PCEDT is finished.

[3] The members that carry the centre of the intonation in the sentence are capitalized (in the translation).

[4] In fact, the topic-focus articulation of the given sentence is the same regardless on the language. However, we operate with a parallel corpus – the English part contains original texts and the Czech one their translations. It is possible that the Czech translations could be inaccurate in some cases – especially regarding the topic-focus articulation. Therefore, the value of contextual boundness could differ in both parts of parallel corpus in a few cases.

So far, the automatic procedure was used for pre-annotation of a sample of the PCEDT Czech part and this pre-annotated sample was subsequently manually annotated by a human annotator. The annotator checked the correctness of the pre-annotation and annotated the rest of the nodes (nodes that had not been pre-annotated). Afterwards, it was evaluated how many changes of the automatic pre-annotation of topic-focus articulation the human annotator had to carry out, i.e. how many mistakes the automatic pre-annotation had made in the data.

It should be noted that the goal of the automatic pre-annotation was to help the human annotators with simple decisions, not to classify every sentence member as contextually bound ('t') or non-bound ('f') element. Our intention was to apply only reliable rules and leave too complex decisions (often depending on the meaning of the text) on the human annotator. We wanted to avoid introducing too many errors in the pre-annotation, as human annotators might be prone to overlooking errors in already annotated nodes and concentrate only (or at least better) on the so far unannotated nodes. For the selection of the pre-annotation steps, we estimated their expected error rates (where possible) based on measurements on the topic-focus annotation in PDT (see the expected error rates of the individual pre-annotation steps below in 4.1). For using a rule, we set the maximum number of expected errors to 10 %.

4.1 Steps of the Pre-Annotation

The following steps have been performed during the automatic pre-annotation. For each step (where possible), we give an estimate of the pre-annotation error (expected error rate, EER), based on the measurement of the phenomenon in the data of Prague Dependency Treebank. The steps have been applied in the presented order. Step 10 takes advantage of the cross-language alignment of words in PCEDT.

1. **Nodes generated** on the tectogrammatical layer **without a counterpart on the analytical layer** (i.e. newly added, but not copied nodes in the tectogrammatical representation) and that do not have functor=RHEM (rhematizer), nor t_lemma=#Forn (part of a phrase in a foreign language), get automatically assigned tfa='t', i.e. contextually bound, (EER: 0). For an example, see Figure 1.[5]

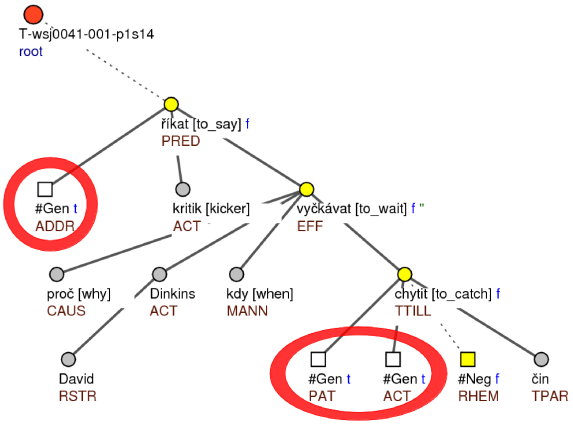

Figure 1: Example of a sentence tree structure in the Czech part of PCEDT: circled nodes represent the automatically pre-annotated sentence members marked as 't' (contextually bound)

Figure 1 represents the following Czech sentence – Example (2) from PCEDT:

(2) *„Proč David Dinkins," říká kritik, „vždycky vyčkává, dokud není chycen při činu?"*

"David Dinkins," says the kicker, "Why does he always wait until he's caught?"

In the surface (analytical) structure of the given sentence with the Czech verb *říkat* (*to say*), the Addressee is not present explicitly although this verb has the Addressee (apart from the Effect, the Actor and the non-obligatory Patient) in its valency frame (*someone*$_{.obligatory\_Actor}$ *says something*$_{.obligatory\_Effect}$ *to someone*$_{.obligatory\_Addressee}$ *about something/somebody*$_{.non-obligatory\_Patient}$). So the Addressee is present only in the deep (tectogrammatical) sentence structure (in Figure 1, it is captured as a small square with the symbol of Addressee ADDR). The sentence members that appear only implicitly in the sentence (as the Addressee in this case) are not supposed to carry some new, important information (because their presence in the /surface part of the/ sentence is not necessary) and therefore they are automatically pre-annotated as contextually bound (fur-

[5] Sentence members (nodes) that are really expressed in the surface sentence structure (that appear on both the analytical and the tectogrammatical layer) are displayed as small circles in the figure. Members that are present only in the deep sentence structure (on the tectogrammatical layer) and do not appear in the surface sentence structure (i.e. not on the analytical layer) are displayed as small squares.

White colour represents contextually bound sentence members (they are also depicted with 't' next to the lemma); yellow colour (light grey in b/w) represents contextually non-bound sentence members (they are depicted with 'f'). The grey members do not have any value of contextual boundness yet (they were not automatically pre-annotated and they will be manually annotated by a human annotator).

ther examples are the sentence members Patient PAT and Actor ACT by the Czech verb *chytit – to catch*: *somebody*$_{\text{.obligatory\_Actor}}$ *catches someone*$_{\text{.obligatory\_Patient}}$, see Figure 1).

2. **Nodes generated at the tectogrammatical layer that are members of coordination/apposition** and have an analytical counterpart (they are copied nodes; it also means that it is not e.g. #Forn), get assigned tfa='t', i.e. contextually bound, (EER: 0), see Example (3) from PCEDT.

(3) *„Nyní," říká Joseph Napolitan, průkopník politické televize, „je cílem jít do útoku jako první, poslední a* **[jít]**$_{t}$ *vždycky."*

"Now," says Joseph Napolitan, a pioneer in political television, "the idea is to attack first and **[to attack]**$_{t}$ *always."*

This pre-annotation step concerns also other cases of sentence members that are not present in the surface (analytical) structure but appear in the deep (tectogrammatical) layer. These nodes are not newly added to the structure, e.g. because of the valency verb frame, but they appeared in some previous structures and they are omitted in the surface structure (and copied to the deep structure) because the reader can understand them easily from the previous context as in the phrases from Example (3): ***to attack*** *first* and (***to attack***) *always*. Since these members (present only implicitly in the sentence) are obviously deducible from the context, they are considered as contextually bound and therefore they are pre-annotated as such.

3. **Nodes where** a grammatical, textual or segment **coreference starts**, get tfa='t', i.e. contextually bound, (EER: 1:100), see Example (4) from PCEDT.

(4) *A Dinkins podle svých*$_{t}$ *slov nevěděl, že* ***muž, kterého***$_{t}$ *platili v rámci kampaně za přesvědčování voličů k účasti, byl odsouzen za únos.*

And, says Mr. Dinkins, he_{t} *didn't know the* ***man his***$_{t}$ *campaign paid for a get-out-the-vote effort had been convicted of kidnapping.*

This step of the automatic pre-annotation takes advantage of the finished annotation of coreference in the PCEDT texts. Sentence elements that are anaphors[6] of a coreference relation are supposed to be contextually bound and therefore they are automatically assigned the value 't'.

There are two coreference relations in Example (4): 1. *Dinkins – svých* (*he*); 2. *muž* (*man*) – *kterého* (*his*). The members that refer to some previous sentence members (*svých* and *kterého* in this case) are automatically pre-annotated as contextually bound.

In another example from PCEDT, depicted in Figure 2, starting nodes (anaphors) of grammatical coreference (three intra-sentential more or less vertical arrows) and textual coreference (two horizontal arrows going from the second tree to the first one) are pre-annotated as contextually bound.

4. Nodes with **functor=PRED** that are **not newly generated** and whose t_lemma does not appear in the previous sentence, get tfa='f', i.e. contextually non-bound, (EER: 1:40), see Example (5) from PCEDT.

(5) *„Pamatujete si na Pinocchia?"* ***říká***$_{f}$ *ženský hlas.*

"Remember Pinocchio?" ***says***$_{f}$ *a female voice.*

The data of previously annotated Prague Dependency Treebank demonstrated that most Predicates (in corpus marked as PRED) are contextually non-bound – therefore, they are pre-annotated as 'f'.

5. **Newly generated** nodes with **functor=PRED** get tfa='t', i.e. contextually bound, (EER: 1:100), see Example (6) from PCEDT.

In contrast to the step 4), Predicates that are not present in the surface sentence structure are pre-annotated as contextually bound, cf. step 3).

(6) *Na obrazovce vidíme dvě zkreslené rozmazané fotografie,* ***pravděpodobně***$_{MOD.f}$ **[vidíme]**$_{t}$ *fotografie dvou politiků.*

The screen shows two distorted, unrecognizable photos, ***presumably***$_{MOD.f}$ **[shows]**$_{t}$ [photos] *of two politicians.*

6. **Other verbal nodes** (gram/sempos=v) **with functor** from the set {ADDR, AIM, CAUS, ACMP, MANN, PAT, EFF, AUTH, BEN, COMPL, EXT, ORIG, RESL, TFHL, TSIN} get tfa='f', i.e. contextually non-bound, (EER: 1:10), see Example (7) from PCEDT.

[6] A reference to an entity or event that has already been mentioned in the preceding text; the two mentions – anaphor (the latter in the text) and antecedent (the former) are connected by a coreference relation.

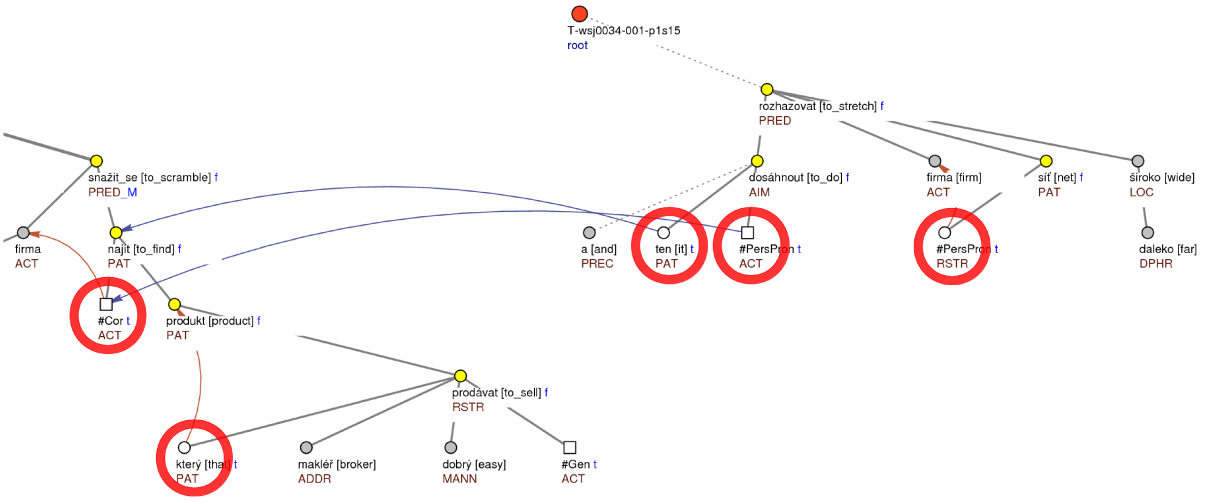

Figure 2: Two trees representing two sentences; the indexes in the sentences and the arrows in the trees denote coreference chains; starting nodes of the coreference links are marked by large circles: *Soukromí investoři se s léty od burzy odvracejí a investiční firmy***[2]** *se snaží* [firmy]**[2]** *najít***[1]** *produkty***[4]**, *které***[4]** *by se makléřům dobře prodávaly. A aby toho***[1]** [firmy]**[2]** *dosáhly, rozhazují firmy***[3]** *své***[3]** *sítě široko daleko.* (In original: *As individual investors have turned away from the stock market over the years, securities firms***[2]** *have scrambled to* [firms]**[2]** *find***[1]** *new products***[4]** *that***[4]** *brokers find easy to sell. And the firms***[3]** *are stretching their***[3]** *nets far and wide to* [firms]**[2]** *do it***[1]**.)

Note that there is no link between *firm* in the first tree and *firm* in the second tree, as only pronominal coreference is annotated in the data. Otherwise, chains [2] and [3] would be one coreference chain.

The example has been cropped to fit the page (the left part of the first tree).

The data of the Prague Dependency Treebank also demonstrated that most sentence members expressed as dependent clauses (i.e. containing a finite verb) and having the semantic role of Addressee, Aim, Cause, Accompaniment, Patient, Effect, Author, Benefactor, Complement, Extent, Origo, Result or Temporal modifications (expressing *for how long* or *since when*) are contextually non-bound – therefore, they are pre-annotated as non-bound also in data of the Prague Czech-English Dependency Treebank.

(7) ***„Porovnejte tyto dva kandidáty na starostu,".Effect$_f$*** *říká hlasatel.*

"Compare two candidates for mayor,".Effect$_f$ *says the announcer.*

7. **Nodes with functor** from the set {PARTL, DENOM, MOD, EXT} get tfa='f', i.e. contextually non-bound, (EER: 1:10), see again Example (6) above from PCEDT.

The data of the Prague Dependency Treebank further demonstrated that most sentence members assigned the semantic role of independent interjectional clause (marked as PARTL), independent non-parenthetical nominal clause (DENOM), atomic expression with a modal meaning (MOD) or adjunct expressing extent (EXT) are contextually non-bound and therefore they are pre-annotated as such.

In the Example (6), the sentence member *pravděpodobně* (*presumably*) is in the role of an atomic expression with a modal meaning (MOD) and therefore it will be automatically assigned the value 'f'.

8. Nodes with **functor=RHEM** (i.e. they have a function of a rhematizer) that are not in the first position in the sentence, get tfa='f', i.e. contextually non-bound, (EER: 1:10), see Example (8) from PCEDT.

(8) *Letošek je rokem, kdy se negativní reklama, po léta přítomná ve většině politických kampaní* ***jen$_f$*** *druhotně, stala hlavní událostí.*

This is the year the negative ad, for years **[only]$_f$** *a secondary presence in most political campaigns, became the main event.*

The rhematizers (as e.g. English particles *only*, *for example*, *also*, *especially*, *principally*) mostly precede a focus element and in the theory of TFA, they are also considered contextually non-bound. However, also contrastive contextually bound expressions can follow the rhematizers – typically at the beginning of the sentence (and in this case, also the rhematizers are contextually bound). Therefore, only such rhematizers are pre-annotated as contextually non-bound that are not placed in the initial position in the sentence.

9. Nodes with **t_lemma=*tady* (*here*)** get tfa='t', i.e. contextually bound, (EER: 1:10), see Example (9) from PCEDT.

Some lemmas (especially with a deictic function like *here*) appear as contextually bound in most cases (but not in all – see e.g. *What happens* ***here***$_{\mathbf{f}}$ *and now?*), which observation is also made use of in the automatic pre-annotation.

(9) *Ředitelka Wardová se rozhodla zbavit se „balastu" v učitelském sboru a obnovit bezpečnost a také* ***tu***$_{\mathbf{t}}$ *byly další nové faktory, které pracovaly v její prospěch.*

Mrs. Ward resolved to clean out "deadwood" in the school's faculty and restore safety, and she also had some new factors **[here]**$_{\mathbf{t}}$ *working in her behalf.*

10. Nodes that are **Czech counterparts of English nodes** that in the English sentence are placed after their governing verb on the surface and that are **preceded by an indefinite article**, get tfa='f', i.e. contextually non-bound, (EER: unknown), see Example (10) from PCEDT.

(10) *The war over federal judicial salaries takes* ***a victim.***↓

Válka o platy federálních soudců si žádá svou první ***oběť***$_{\mathbf{f}}$***.***

In Example (10), the sentence member *victim* is modified by the indefinite article *a* in the English variant of the sentence, which leads to the assumption that this member is contextually non-bound. Since the value of the same sentence member should be identical both in English and in Czech variant of the sentence, also the Czech member *oběť* (that is the counterpart of the *victim*) is supposed to be contextually non-bound.

The following steps of the automatic pre-annotation are performed after the previous steps have been applied on all nodes of the given tree:

11. **Daughters of a verb that has tfa='f'** and that is not on the first or second position (in its clause), if they appear after the governing verb on the surface, get tfa='f', i.e. contextually non-bound, (EER: unknown), see Example (11) from PCEDT.

(11) *Na konci druhé světové války se Německo* ***vzdalo***$_{\mathbf{f}}$ ***dříve než Japonsko***$_{\mathbf{f}}$***...***

At the end of World War II, Germany ***surrendered***$_{\mathbf{f}}$ ***before Japan***$_{\mathbf{f}}$***...***

This step of the pre-annotation makes use of the fact that in Czech, the surface word order often is used to express the topic-focus articulation. Under the condition that the contextually non-bound predicative verb is placed further to the right than on the second position in the sentence and that the sentence has a non-marked word order[7] (i.e. emotionally neutral), it is possible to assume that the sentence members following the predicative verb are contextually non-bound.

12. Nodes with **functor=RSTR** that are **daughters of a node with tfa='f'**, get tfa='f', i.e. contextually non-bound, (EER: 1:30).

(12) *Zasedání společného výboru sněmovny a senátu se koná v případě, že sněmovna a senát schválí zákon v* ***odlišné***$_{\mathbf{f}}$ *podobě.*

The Senate-House conference committee is used when a bill is passed by the House and Senate in ***different***$_{\mathbf{f}}$ *forms.*

The final step of the automatic pre-annotation is based on the fact that the adnominal adjuncts modifying its governing noun (in the annotated corpus marked as RSTR) often have a very high degree of communicative dynamism because their primary function is to specify something. Therefore, they are pre-annotated as contextually bound (if they modify a non-bound element at the same time).

5 Evaluation of the Automatic Pre-Annotation

At the time of submitting the final version of the paper, more than one thousand automatically pre-annotated sentences have also been manually annotated by a human annotator[8] and could be used for evaluation of the pre-annotation.

In 59 documents (1,145 sentences, 22,436 nodes on the tectogrammatical layer), 7,864 nodes out of 19,105 tfa-relevant nodes have been automatically pre-annotated (i.e. 41.1 %).

Table 1 gives an overview of how many times the individual pre-annotation steps have been applied. Based on the estimates presented in Sec-

[7] The human annotator decides whether the word order is marked or non-marked (it is not possible to check it automatically in our procedure of pre-annotation).

[8] There were actually two annotators, working on different parts of the data. For simplicity, we refer to them as 'a human annotator'. Only during a training phase (performed on a few documents), the two annotators worked on the same data and their discrepancies were subsequently checked by an arbiter and discussed.

tion 4.1 (for the two unknown estimates in steps 10 and 11 we used EER: 1:10), we can calculate the expected number of errors in the pre-annotation as (about) 340 errors.

step	short description	count
1	generated, no a-counterpart	1,988
2	generated, member of coord/app	127
3	anaphor of a coreference	742
4	PRED, not generated	1,189
5	PRED, generated	0
6	other verbal nodes (set of func.)	825
7	set of functors	435
8	RHEM (not first in sentence)	366
9	t_lemma=*tady* (*here*)	8
10	indefinite article in English	779
11	subseq. daughter of a verb in focus	237
12	RSTR daughters of a node in focus	1,168

Table 1: Usage of the individual pre-annotation steps

In the manual annotation, the annotator changed the pre-annotated value in 294 cases (i.e. 3.7 % of pre-annotated nodes). Table 2 shows details on the manually performed changes.

	pre-annotated value	
	't'	'f'
changed to 'c'	11	26
changed to 't'	-	244
changed to 'f'	13	-
no change	2,841	4,729

Table 2: The distribution of changes of automatically pre-annotated TFA-values manually made by human annotators

The numbers show that the automatic pre-annotation is more successful in marking contextually bound sentence members, as only 0.8 % of nodes pre-annotated as 't' and 5.4 % of nodes pre-annotated as 'f' were manually changed to another value.

	PDT 2.0	sample of PCEDT
contr. contextually bound ('c')	5.4 %	5.7 %
non-contr. contextually bound ('t')	31.3 %	33.6 %
contextually non-bound ('f')	63.3 %	60.7 %

Table 3: The percentage distribution of manually annotated TFA-values in PDT (training data) and so far annotated sample of the Czech part of PCEDT

The inability of the pre-annotation procedure to set the 'c' value (contrastive contextually bound) does not harm the results much, as only 37 (0.5 %) pre-annotated nodes were manually changed to this value, and the overall ratio of contrastive contextually bound nodes among all (manually) annotated nodes both in PDT and PCEDT is less than 6 % (see Table 3).

The main limitations of the pre-annotation are in its coverage (more than half of the nodes are not pre-annotated) and in its natural inability to take the meaning of the text into account (and thus being unable to better distinguish between 't' and 'f' values).

From another point of view, the results suggest that the expected error rates (estimated on PDT) are accurate and that the automatic pre-annotation is sufficiently reliable and serves as a substantial help to the annotators.[9]

6 Conclusion

The paper presented the first part of the project of parallel annotation of topic-focus articulation in the Prague Czech-English Dependency Treebank (PCEDT). We described the annotation principles and schemes, and elaborated on 12 automatic steps of the pre-annotation procedure for the Czech part of the treebank. The pre-annotation is able to mark over 40 % of the whole text (the rest is supposed to be annotated by human annotators). It can distinguish between contextually bound and non-bound sentence elements with the average success rate over 96 %, as shown by the evaluation on manually annotated texts.

Acknowledgment

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875). This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

E. Bejček, J. Panevová, J. Popelka, P. Straňák, M. Ševčíková, J. Štěpánek, Z. Žabokrtský. 2012. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th Inter-*

[9] Of course, it is a matter of discussion (and testing), how much effort of the human annotator such a pre-annotation saves and how to set the reliability limit for the rule selection.

national Conference on Computational Linguistics (Coling 2012), Mumbai, India, pp. 231–246.

S. Calhoun, M. Nissim, M. Steedman, J. Brenier. 2005. A Framework for Annotating Information Structure in Discourse. In: *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 45–52. URL http://aclweb.org/anthology/W/W05/W05-0307.

P. Cook, F. Bildhauer. 2011. Annotating information structure. The case of "topic". In: S. Dipper & H. Zinsmeister (eds.), *Beyond Semantics.Corpus based Investigations of Pragmatic and Discourse Phenomena,* Ruhr Universität Bochum, Bochumer Linguistische Arbeitsberichte, pp. 45–56. URL http://www.linguistics.ruhr-uni-bochum.de/bla/beyondsem2011/cook_final.pdf.

S. Dipper, M. Götze, S. Skopeteas (eds.). 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure,* vol. 7 of Interdisciplinary Studies on Information Structure. Potsdam, Germany: Universitätsverlag Potsdam. URL http://www.sfb632.uni-potsdam.de/publications/isis07.pdf.

S. Dipper, M. Götze, M. Stede, T. Wegst. 2004. ANNIS: A Linguistic Database for Exploring Information Structure. In Ishihara, S., Schmitz, M., Schwarz, A. (Eds.), *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS) 1,* pp. 245–279. Potsdam: University publishing house Potsdam.

K. Donhauser. 2007. Zur informationsstrukturellen Annotation sprachhistorischer Texten. *Sprache und Informationsverarbeitung 31,* pp. 39–45. URL http://www.sfb632.uni-potsdam.de/publications/B4/donhauser_2007.pdf.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková-Razímová. 2006. *Prague Dependency Treebank 2.0.* Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, http://www.ldc.upenn.edu, Jul 2006.

J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, Z. Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 3153–3160.

E. Hajičová, J. Havelka, K. Veselá. 2005. Corpus Evidence of Contextual Boundness and Focus. In: *Proceedings of the Corpus Linguistics Conference Series*, University of Birmingham, Birmingham, UK, ISSN 1747-9398.

E. Hajičová, J. Panevová, P. Sgall. 2000. A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank. *Technical report tr-2000-09,* ÚFAL/CKL. URL http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/tmanual/tmanen.pdf. In cooperation with A. Böhmová, M. Ceplová and V. Řezníčková. Translated by Z. Kirschner, E. Hajičová and P. Sgall.

E. Hajičová, B. H. Partee, P. Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content.* Dordrecht, Boston: Kluwer Academic Publishers.

M. P. Marcus, B. Santorini, M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19(2)*, pp. 313–330.

M. Mikulová et al. 2005. *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. The Annotation Guidelines*. Prague: UFAL MFF. Available at: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

M. Nissim, S. Dingare, J. Carletta, M. Steedman. 2004. An annotation scheme for information status in dialogue. In: *Proceedings of the 4thConference on Language Resources and Evaluation.* Lisbon, Portugal. URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/638.pdf.

P. Paggio. Annotating Information Structure in a Corpus of Spoken Danish. 2006. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 1606–1609, Genoa, Italy.

O. Postolache. 2005. Learning Information Structure in The Prague Treebank. In: *Proceedings of the ACL Student Research Workshop*, pp. 115–120, Ann Arbor, Michigan, June 2005.

P. Sgall. 1967. *Generative description of language and the Czech Declension (in Czech)*. Prague: Academia.

P. Sgall, E. Hajičová, E. Benešová. 1973. *Topic, focus and generative semantics (Vol. 1).* Kronberg Taunus: Scriptor Verlag.

P. Sgall, E. Hajičová, J. Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer.