Topic-focus articulation and anaphoric relations: corpus based probe

Lucie Kučová, Kateřina Veselá, Eva Hajičová, and Jiří Havelka

1. The objective of the paper is to analyze certain interrelationships between the information structure, i.e. the topic-focus articulation (TFA) of sentences, and anaphoric relations, on the material achieved during the annotation of TFA and of coreference in the Prague Dependency Treebank (PDT).

2. Underlying layer of annotation of PDT

2.1 Prague Dependency Treebank (PDT) is conceived of as a collection of 3,168 samples of continuous running Czech texts (taken at random from the Czech National Corpus), annotated – besides a complex scheme of morphemic tags – on two layers of dependency-based sentence structure, the first of which – the analytic one – is considered to be an intermediate step towards the underlying level of annotation, the so-called tectogrammatical tree structures (TGTSs), in which nodes are also reconstructed for items deleted in the surface shape of the sentences. These structures are designed in a way that allows i.a. for an inclusion of information on both intra- and inter-sentential coreference relations.

2.2 In addition to the deep syntactic dependency relations in the tree structure, individual nodes are assigned one of the three values of **contextual boundness**: non-contrastive contextually bound “t”, contrastive contextually bound “c” and contextually non-bound “f”. This information at individual nodes of the dependency tree structure makes it possible to derive the division of the sentence into **topic** (in the prototypical case: what the sentence is about) and **focus** (what the sentence says about the topic); the basic algorithm for this procedure was formulated by Hajičová and Sgall (see Hajičová and Sgall 1985) and its implementation and testing on PDT is reported in Hajičová, Havelka and Veselá (2005).

2.3 In a separate path through the corpus annotated on this underlying level, basic **coreference relations** are being marked independently of the TFA values. In our project, two types of coreference are distinguished: **grammatical** – with verbs (and also some nouns) of control, with reflexive pronouns, with verbal complements and with relative pronouns – and **textual**, which may cross sentence boundaries. Both endophoric (anaphora) and exophoric (deixis) relations are taken into account as well as cataphora (see Kučová and Hajičová 2004).

For the annotation of grammatical coreference (which has been given a systematic account in the description, see Kučová et al. 2003) a semi-automatic procedure has already been implemented, which is giving rather encouraging results.

The manual annotation of textual coreference is carried out with the use of a user-friendly tool in the TrEd editor used for tree-structure assignment (Kučová et al. 2003). The use of such an original user-friendly software tool results in more accurate and consistent annotations and speeds up the whole process. It also makes it possible to apply annotation on relatively large corpus data (in our case, the procedures described above have already been applied to the whole set of 50,000 sentences annotated on the underlying syntactic level). Some steps already undertaken in this direction (involving only the resolution of textual coreference links „starting“ with the tectogrammatical lemma PersPron, which stands for personal and personal possessive pronouns[1]) brought encouraging results – the succes rate is 60.4%.[2]

For the time being, we concentrate on cases of textual coreference in which demonstrative or anaphoric pronouns (also in their zero form in the surface shape of the sentence) are used. The following types of textual links are distinguished:

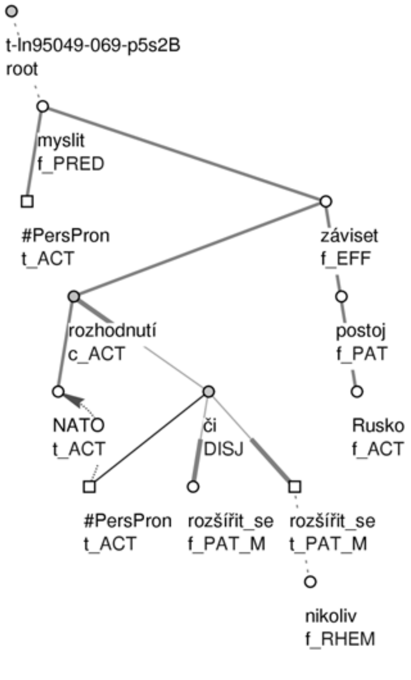
[1] Be they expressed on the surface or restored during the annotation of the tectogrammatical tree structure.

[2] The resolution system is described in Kučová and Žabokrtský (2005).

(a) a link to a particular node if this node represents a referent (antecedent) of the anaphor;
(b) a link to the governing node of a subtree if the antecedent is represented by this node plus (some of) its dependents; this is also the way how a link to a previous/following clause or a whole sentence is being established;
(c) a specifically marked link (Segm for 'segment') denoting that the antecedent is a whole segment of (previous) text larger than one sentence or phrase, including also the cases in which the antecedent is understood by inferencing from a broader co-text;
(d) a specifically marked link (Exoph for 'exophor') denoting that the referent is 'out' of the co-text and is known only from the situation.

3. At present, we have at our disposal both the TFA annotation and an indication of the coreference relations (at least for a limited but precisely specified group of anaphors, see above) of 22,889 nodes of tectogrammatical sentence structures. This amount of data allows us to ask several questions on the interrelationships of the two aspects.

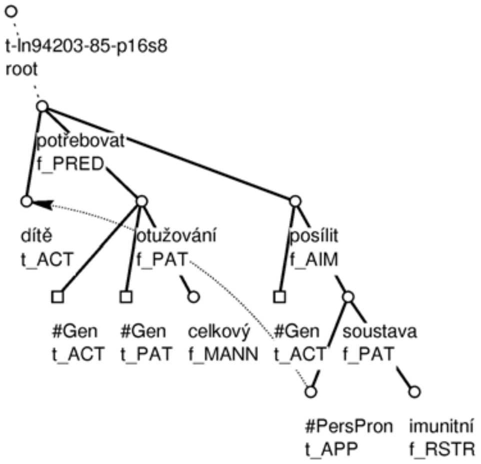
3.1 One of the first questions that come to mind is whether it is always so that a contextually bound item refers anaphorically (including the exophoric reference). As at this stage of annotation we annotate only those nodes for textual coreference of personal and possessive pronouns of the 3rd person singular and plural and demonstrative pronouns, including cases where a pronoun of this category is deleted in the surface shape of the sentences, see ex. (1), this is a trivial question and no counterexamples have been found (except for evident annotators' mistakes).



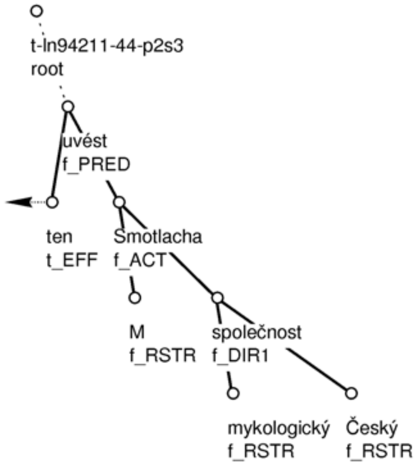
(1) *Myslíte, že rozhodnutí <u>NATO</u>, zda se **[ono]** rozšíří, či nikoli, bude záviset na postoji Ruska?*
[Do-you-think that the-decision of-<u>NATO</u> whether Refl. **[it]** will enlarge or not will depend on the-attitude of-Russia?][3]

[3] The English versions of the Czech example sentences are just word-for-word translations (the context need not be translated word for word). The anaphoric items relevant for our discussion are printed in bold. The antecedents of the anaphoric items are underlined. The figures with the tectogrammatical tree structures show the example sentences as they are annotated in PDT 2.0 (but only the relevant part of the whole annotation is shown). For a node, the first line gives its so-called tectogrammatical lemma (there are several special lexical values, e.g. Gen stands for a General Participant). In the second line you can find the value of contextual boundness (t, c, or f), the labels in capital letters are abbreviations for the valency values (functors; ACT stands for Actor, PAT for Patient, etc., RHEM denotes the function of Rhematizer (focalizer)), nodes in coordination are marked by the suffix M. Reconstructed nodes have the shape of a rectangle rather

3.2 It is not surprising that most of the anaphoric links (22,582, i.e. 98.6%, out of which 21,990 refer to a particular node, 494 refer to a segment and 98 are exophoric) lead from nodes annotated as "t" (non-contrastive contextually bound). As noted above, the antecedent may be a single node to which the link points or a whole subtree with a link pointing to the governor of the given subtree (see examples (2) and (3), respectively), a segment (4), or an exophoric reference (5).

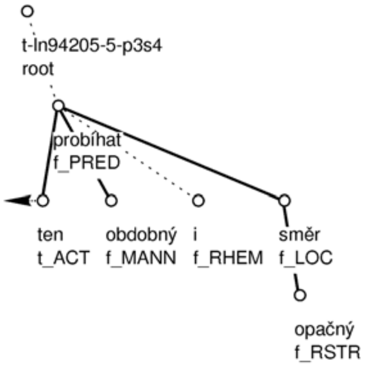


(2) *Dítě potřebuje otužování, aby byla posílena* ***jeho*** *imunitní soustava.*
[Child needs hardening so-that was strengthened **its** immunity system.]

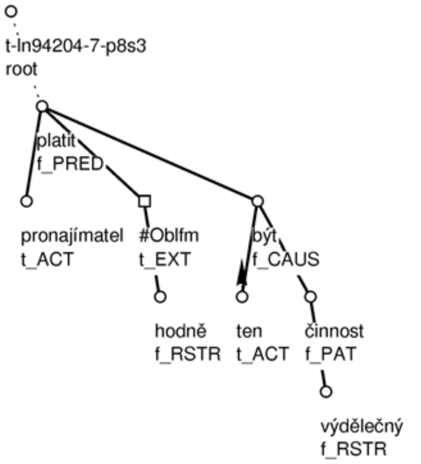


(3) (*Začaly růst i houby jedovaté.*)
Uvedl **to** *M. Smotlacha z České mykologické společnosti.*
[(Also the poisonous mushrooms started to grow.)
Adduced **that** M. Smotlacha from Czech mycological association.]

than a circle, the different shapes of edges only help to indicate some aspects of the annotation. (The roots of the trees are auxiliary technical nodes for which the identifiers of the sentences are given.)



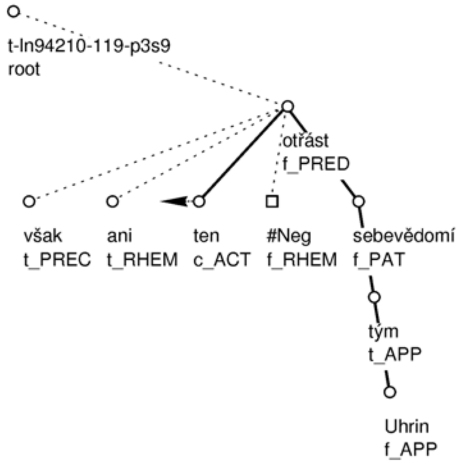
(4) *(Mascu pro Windows, program liberecké firmy Merz, zvládne každý, kdo (...) ovládá Windows. Za 2200 Kč (bez DPH) získáte již zmíněný program a také kabel pro spojení Casia s PC. Musíte je jen vzájemně propojit, spustit Mascu, označit v diáři to, co si přejete přenést, a po chvíli se vám data objeví v počítači.)*
Obdobně ***to*** *probíhá i v opačném směru.*
[(Everyone who is good at Windows is able to work with Masca, a program by Merz, a company from Liberec. You can get the referred program and also the cable for a connection of Casio with the PC for 2000 CZK (VAT excluded). You must only interconnect them, initiate Masca, mark what you want to transfer in your diary, and in a while you can see your data in your PC.)
Similarly **that** proceeds also in opposite direction.]



(5) *Pronajímatel platí víc, protože* ***to*** *je výdělečná činnost.*
[Landlord pays more because **that** is profitable activity.]

3.3 A less trivial question, possibly throwing some light on the not yet fully understood phenomenon of "contrastive topic," is the following: what are the most typical anaphoric links from a node assigned the value "c"?

There are 128 anaphoric links from nodes marked as a **contrastive** (part of) **topic** (TFA value = c). Most links (121) point to a particular node in the preceding co-text; it can be well understood that no cases have been found in which the link going from a contrastive item is exophoric, and just 7 lead to a previous segment (cf. ex. (6)).



(6) *(Hosté se dostali do vážnější akce po čtyřiadvaceti minutách.*
Nedvědovi utekl Vella, odcentroval a Laverla mířil těsně vedle.)
Ani ***to*** *však neotřáslo sebevědomím Uhrinova týmu.*
[(Guests got to a promising action after 24 minutes.
Vella fled Nedvěd, he passed the ball and Laverla nearly missed.)
Even **that** however did-not-shake self-consciousness of-Uhrin's team.]

3.4 The set of examples in which a coreferential link points to a particular node may be further subdivided into several groups, according to the "scope" of the antecedent: it may be either a single item (ex. (7)), or a whole subtree the governor of which is the node to which the link leads (ex. (8)); the latter case covers also instances where the antecedent is a whole sentence (the link leads to the main verb):

(7) *Zpravidla jsou na nich novinky a o* ***ty*** *právě zákazníci stojí.*
[As-a-rule are on them hot-news and about **those** exactly customers care.]

(8) *V parlamentu jsou sice poměrně početné skupiny zaměřené proti vládě a premiérovi, avšak i* ***ony*** *jsou si vědomy toho, že…*
[In Parliament are though relatively numerous groups directed against government and Prime-Minister, but even **they** are Refl. aware of-the-fact that…]

3.5 Further interesting cases from the point of view of the relationships between topic-focus articulation and coreference are those where the coreferential link leads from a node with the TFA value f (i.e. from a contextually non-bound node). In our corpus, we have found 179 anaphoric links leading from nodes marked as f, with the following distribution: 155 lead to a particular node, 2 refer to a segment, and 22 relations are exophoric. The examples of exophoric relations are rather obscure – even in such an example as (9) the exophoric interpretation is not clear:

(9) *Následuje dramatická pauza a pak již vchází* ***On*** *nebo* ***Ona****.*
[There-follows dramatic pause and then already enters **He** or **She**.]

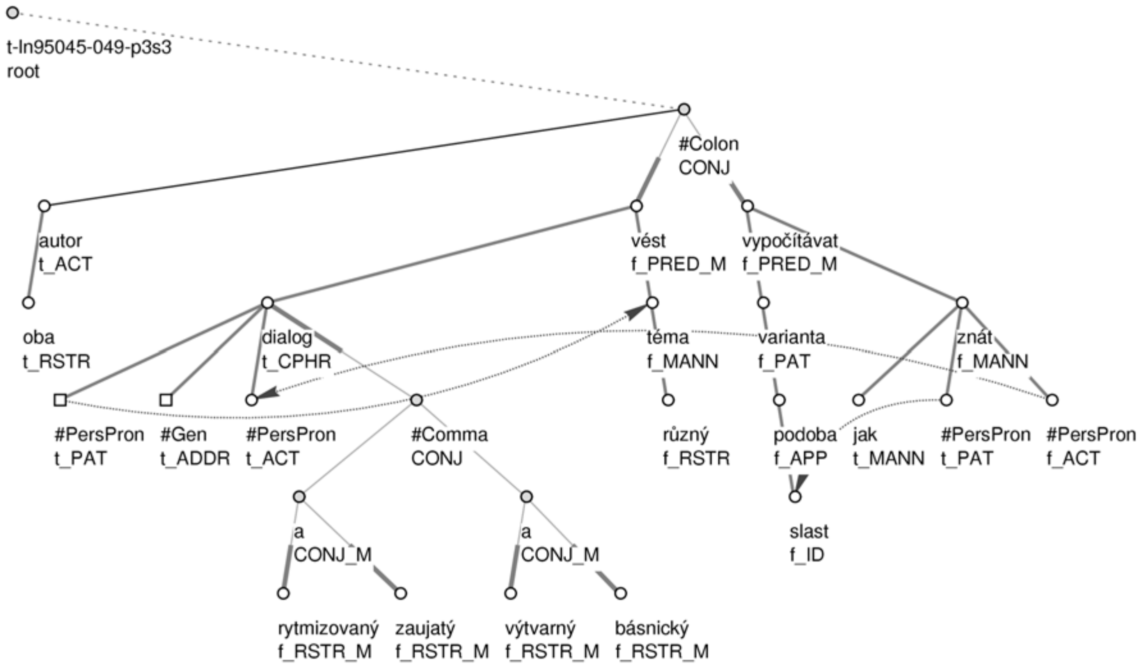
Some examples are phraseological constructions, which eventually would be represented as a single node with no anaphoric value (cf. ex. (10)):

(10) *(...) nemáme dost peněz na* ***to*** *či* ***ono****.*
[(…) we-have-not enough money for **this** or **that**.]

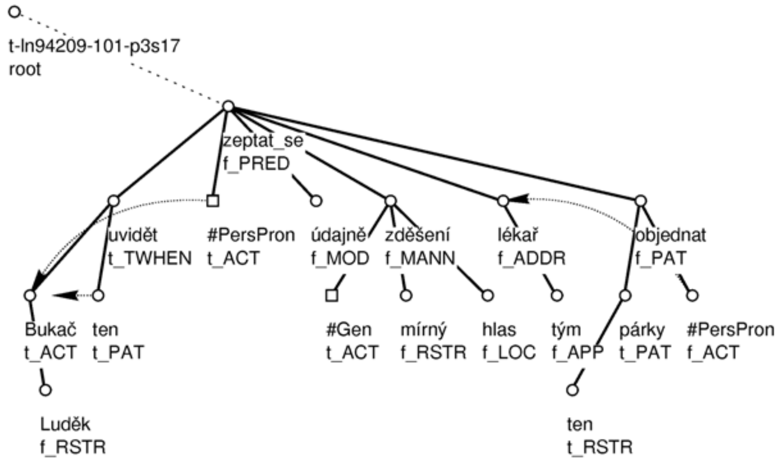
As sentence (11) demonstrates, a personal pronoun can be also used as a metalinguistic expression:

(11) *Připisuje se jakémusi záhadnému* ***ono****.*
[It-is-ascribed Refl. to-some mysterious **it**.]

However, cases in which the anaphoric link leads from a contextually non-bound node to an antecedent in the previous context are not rare (be it a single node or a subtree, see (12) and (13), respectively) and their existence confirms that it is not correct to identify the linguistic information structure of the sentence directly with the cognitive given-new distinction.



(12) *Svůj (...) dialog vedou oba autoři na různá témata:* *[oni]* *vypočítávají varianty podob slastí tak, jak ji znají* ***oni.***
[Their (…) dialogue lead both authors on various topics: [they] enumerate variants of-forms of-delights so as it know **they**.]



(13) *(...) zeptal se údajně s mírným zděšením v hlase* *lékaře týmu**, zda ty párky objednal* ***on.***
[(…) he-asked Refl. allegedly with slight panic in voice doctor of-team whether those sausages ordered **he**.]

3.6 The anaphoric reference to a segment needs some more specific delimitation. It is our future task to examine whether some conditions of such a delimitation can be based on the TFA annotation of the sentences included in the segment referred to.

4. Summary

We are well aware that the data collected up to now as for the two aspects – information structure and coreference – are rather sparse and need a completion and further examination. However, the probe we describe in the paper confirms that if the coreference assignment is not done selectively but if it is an integral part of a large scale annotation of underlying sentence structure (along with the annotation of the information structure of sentences), a corpus annotated in this way prepares solid grounds for further linguistic investigations of discourse patterns.

5. Acknowledgements

The research reported on in this paper has been supported by the grant of the Charles University in Prague 207-10/203329.

References

Hajičová Eva and Petr Sgall (1985), Towards an Automatic Identification of Topic and Focus. In: Proceedings of the European Chapter of the Association for Computational Linguistics, Geneva, pp. 263–267.

Hajičová Eva, Jiří Havelka, and Kateřina Veselá (2005), Corpus evidence of contextual boundness and focus. Accepted for the conference Corpus Linguistics 2005, Birmingham.

Kučová Lucie, Veronika Kolářová, Petr Pajas, Zdeněk Žabokrtský, and Oliver Čulo (2003), Anotování koreference v Pražském závislostním korpusu [Annotation of coreference in the Prague Dependency Treebank]. Technical Report TR-2003-19 of the Center for Computational Linguistics, Charles University, Prague.

Kučová Lucie and Eva Hajičová (2004), Coreferential Relations in the Prague Dependency Treebank. Paper presented at the DAARC 2004, San Miguel, Azores.

Kučová Lucie and Zdeněk Žabokrtský (2005), Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution. Accepted for the conference TSD 2005, Karlovy Vary.

Prague Dependency Treebank 2.0, to be published by Linguistic Data Consortium in 2005, `http://ufal.mff.cuni.cz/pdt2.0/`.