# Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Coreferential Mark-Up

Lucie Kučová, Eva Hajičová

## 1. Introduction

The complex annotation scenario of the Prague Dependency Treebank (henceforth PDT, i.e. a collection of 2000 samples each containing 50 continuous sentences from running Czech texts; the samples are taken at random from the Czech National Corpus), is conceived of as an annotation consisting of three layers, namely the morphemic (POS tagging taking into account the rich inflecting inventory of characteristics of word forms), the analytic (reflecting the surface shape of the sentences) and the tectogrammatical (capturing the underlying syntactic relations). The tree structures on both the analytic and the tectogrammatical level are dependency trees. Attention has been always focused, however, on the tectogrammatical structures (abbreviated henceforth as TGTS), the analytic ones being understood as a kind of an intermediate stage that has no theoretical status, although it might help to formulate automatic procedures for a transition from the surface shape of the sentences to their underlying representations. The specification of the shape of the TGTSs is based on an explicit, formal linguistic framework developed by the Prague team of theoretical and computational linguistics since the late sixties (Functional Generative Description, Sgall et al. 1986); at the same time, the application of the annotation to "real" language helps to discover new subtleties and thus has consequences for the formal description.

The annotation on the morphemic level is carried out by a stochastic tagger (see Hajič and Hladká 1998) based on a detailed computational morphological analysis of Czech (see Hajič 2004); the annotation on the analytic layer is performed semi-automatically, with the use of a dependency-based modification of Collins' parser (Collins 1999) which cuts down the manual tree-editing to about 20% of the whole work.

The tectogrammatical annotation is also semi-automatic, though the load of the manual work is much heavier than with the annotation of the analytic level. The human annotators have as their input analytic tree structures preprocessed and modified by an automatic procedure deleting the function words (such as prepositions, subordinate conjunctions and modal verbs) and adding their values to the autosemantic nodes of the tree as well as making some further adjustments that can be done automatically. The annotators are helped by an extremely user-friendly tree editor (see TRED) and by several other useful tools such as two valency dictionaries (one, so-called PDT-VALLEX, which is being compiled "on the way", that is which helps the annotators to preserve consistency in the assignments of valency roles [see Hajič and Urešová 2003], and VALLEX1.0, which is compiled "top-down", i.e. Czech verbs of a certain frequency or type are selected and analysed in detail as for their valency characteristics, combimatorial features etc. [Lopatková et al. 2003]).

The (mostly manual) annotation of the tectogrammatical level proceeds basically in three steps or phases: first, the underlying syntactic tree structures are established (or, more precisely, the input analytic tree structures are manually modified and labeled in order to obtain the tectogrammatical tree structures, including the addition of nodes that are deleted on the shallow structure of the sentences and the mark-up of cases of grammatical coreference relations; for the distinction between grammatical and textual coreference, see … and below in Sect. 2). These structures are the input for a group of annotators who – in the second phase – add to the labels of the nodes one of the three values of the topic-focus (TFA) attribute (see Hajičová et al. 2003); the trees with this assignment will serve as an input for an automatic procedure of the bipartition of the sentence into topic and focus formulated on the basis of the definition of focus and topic (see Sgall 1979). In the third phase, another group of annotators processes

(again with help of a very useful user-friendly editor) the tectogrammatical tree structures and adds coreferential links to nodes that stand for a (possibly zero) personal or a demonstrative pronoun. 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 (with the success rate for some phenomena concerned reaching almost 97 %).

The present paper is devoted to the annotation of textual coreference links, with some introductory remarks on the two types of coreference in general (Sect. 2). The annotation scheme is described in Sect 3, with some statistics presented and discussed in Sect. 3.3. In Sect. 4 some problematic issues and open questions are being listed illustrated by examples from the PDT.

## 2. Two types of coreference

In our project, two types of coreference are distinguished: grammatical coreference (typically within a single sentence) and textual coreference (which may but need not cross sentence boundaries); the latter type of coreference covers both both endophoric and exophoric links.

The grammatical coreference involves verbs of control, reflexive pronouns, verbal complements, reciprocity and relative pronouns. In the annotation scheme, the kinds of grammatical coreference are encoded by different lexical values of the node labels; e.g. the "reconstructed" node for the subject of the embedded infinitival clause with verbs of control, such as *slíbit* (to promise somebody to do something), *přesvědčit* (to convince somebody to do something), *požádat* (to ask somebody to do something), carries a label Cor.

In the present stage of annotation of textual coreference, we restrict ourselves to cases of textual coreference in which a demonstrative or an anaphoric pronoun (also in its zero form) are used (with the demonstrative pronoun, we consider only its use as a noun, not as an adjective). We do not include cases of exophoric coreference rendered by a pronoun of the 1st and 2nd persons (be they expressed explicitly or by a zero form, i.e. deleted in the surface shape of the sentence). For the purpose of the present paper, we also leave out of consideration cataphoric reference such as in (1)

(1)*„Vidím ho.“ Velitel: „Oddělej ho.“ Čečen se hroutí.*

(“I see him.” Commander: “Kill him.” [The] Chechen falls down.)

For the time being, we also do not cover the so-called bridging anaphor though some preparatory steps in this direction have already been undertaken.

## 3. Annotation scheme

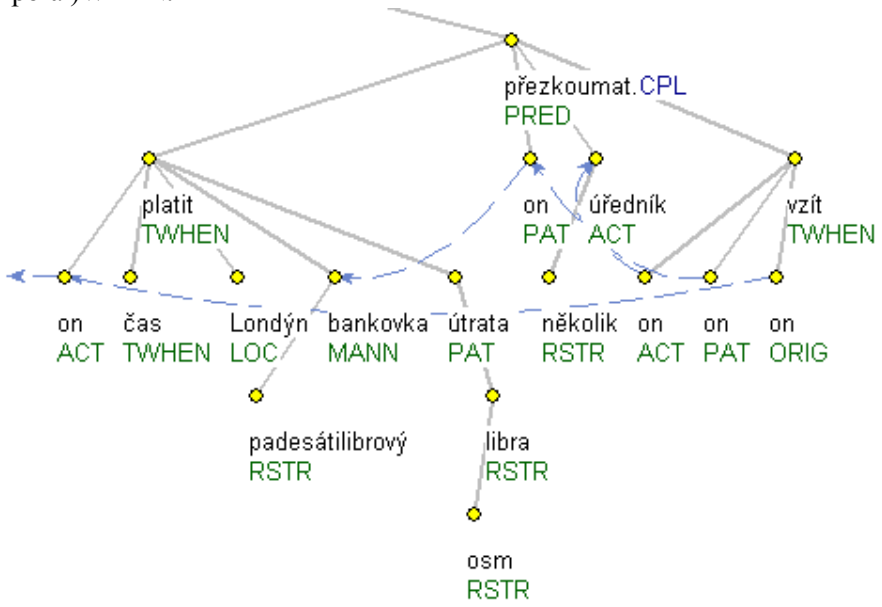
**3.1** In the Prague Dependency Treebank, coreference is understood as an asymmetrical binary relation between nodes of a TGTS (not necessarily the same TGTS), or, as the case may be, as a relation between a node and an entity that has no corresponding counterpart in the TGTS(s). The node from which the coreferential link leads, is called an anaphor, and the node, to which the link leads, is called an antecedent.

The present scenario of the PDT provides three coreferential attributes: coref, cortype and corlemma. The attribute coref contains the identifier of the antecedent; if there are more than a single antecedent of one anaphor, the attribute coref includes a sequence of identifiers of the relevant antecedents. The attribute cortype includes the information on the type of coreference (the possible values are gram for grammatical and text for textual coreference), or a sequence of the types of coreference, where each element of type corresponds to an element of coref. The attribute corlemma is used for cases of a coreference between a node and an entity that has no corresponding counterpart in the TGTS(s): for the time being, there are two possible values of this attribute, namely *segm* in case of a coreferential link to a whole segment of the preceding text (not just a sentence) and *exoph* in case of exophoric relation.

Since in the present shape of the TGTS's every node of a TGTS has an identifier of its own, it is no longer necessary to preserve the original scheme of coreferential attributes the values of which explicitly copied the lemma, the functor or grammateme and the position of the antecedent (see Hajičová, et al. 2000). The identifier present in the attribute coref uniquely points to the antecedent(s) and it is a simple programming task to select the specific information on the antecedent.

In order to facilitate the task of the annotators and to make the resulting structures more transparent and telling, the coreference relations are captured by arrows leading from the anaphor to the antecedent and the types of coreference are distinguished by different colours of the arrows. There are certain notational devices used in cases in which the antecedent is not within the co-text (exophoric coreference), or when the link should lead to a whole segment rather than to a particular node. If the anaphor corefers to more than a single node or to a subtree, the link leads to the closest preceding coreferring node (subtree). If there is a possibility to choose between a link to an antecedent or a postcedent, the link always leads to the antecedent.

In Fig. 1 we present an example of coreference assignment by means of links used by the annotators; sentence (2) is taken from the PDT (the identification number of the sentence is given in the brackets). The following abbreviations are used as the labels for the valency relations (functors): PRED(icate) for the main verb, ACT(or), PAT(ient), LOC(ation), R(e)STR(ictive attribute), ORIG(in), T(emporal)WHEN.




(2) *Když před časem platil v Londýně padesátilibrovou bankovkou útratu osm liber, přezkoumalo ji několik úředníků, než ji od něho vzali.* (lk4#26)
(Lit.: When before time he-paid in London [with] fifty pound banknote amount [of] eight pounds, checked it several clerks[-subj] before it from him they-took.
E.: When he paid a sum of eight pounds with a fifty pound banknote in London some time ago, several clerks have checked it before they took it from him.)

**3.2** To summarize, at the present stage, the following types of textual coreference links are distinguished (some issues related to these types are discussed in Sect. 4 below):

(a) a link to a particular node if this node represents an antecedent of the anaphor (in ex. 3, the link from *ono* leads to *NATO)*:

(3) *Myslíte, že rozhodnutí NATO, zda se **[ono]** rozšíří, či nikoli, bude záviset na postoji Ruska?*
(Do you think that the decision of NATO whether **[it]** will be enlarged or not will depend on the attitude of Russia?)

(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 (ex. 4) or a whole previous sentence (ex. 5) is being established; in (4) the link from *tím* ([by] this) points to the root of the tree (*vynesou*, elevate), i.e. to the main verb of the second conjunct, in (5) the link from *toho* (this) points to the governing verb of the whole sentence (*připravuje*, prepares):

(4) *Ale je něco jiného, když je někdo podnikatel a pak jde do politiky, anebo jestli někoho politické změny vynesou na špičku a on **toho** pak využívá k hospodářské činnosti a zastává vysoké funkce ve velkých firmách.*
[But it is a different thing when someone is an enterpreneur and then goes into politics than when political changes elevate somebody to the top and he then uses **this** in his economic activities and attains a high position in a big firm.]

(5) *Generál kromě toho připravuje nařízení, podle něhož se na něj budou moci obrátit všichni, kteří se domnívají, že se jim děje bezpráví. Hodlá **tím** předejít tomu, aby se redukce armády stala záminkou k vyřizování účtů.*
(The general also prepares an order according to which all who think that harm is being done to them can turn to him. By **this** he intends to avoid a reduction of the army being a pretext for paying off old scores.)

(c) a specifically marked link (*segm*) denoting that the referent is a whole segment of (previous) text larger than one sentence, or phrase, including also the cases, when the antecedent is understood by inferencing from a broader co-text (ex. 6 and 7):

(6) *Podle Kohla nelze zapomenout na to, že Německo přepadlo 22. června 1941 Sovětský svaz. Němci jménem Německa přivodili ruskému lidu nesmírné utrpení. Stejně tak nelze zapomenout, co Rusové způsobili Němcům. Z **toho** všeho si chceme vzít společné poučení.*
(According to Kohl it should not be forgotten that on June 22, 1941 Germany attacked the Soviet Union. Germans on behalf of Germany caused the Russians to suffer immensely. It also cannot be forgotten what the Russians did to Germans. From all **this** we should learn.)

(7) *Potentáti v bance koupí za deset, prodají si za patnáct. Ale povede to k rychlému přerodu. Zmizí výměry kolem 25 ha, přibude vlastníků kolem 500. Odhaduji, že do dvou let budou schopni splatit bance dluh a třetím rokem už budou dělat na sebe. A na práci najmou jen schopné lidi, bude to v jejich zájmu. Kdo **to** pochopí, má náskok.*
(The big shots buy in a bank for ten and sell for fifteen. But this leads to a rapid transformation. The acrages of about 25 ha disappear, the number of owners raises to 500. I guess that within two years they will be able to pay back the debt to the bank and in the third year they will work for themselves. And they will hire only capable people, it will be in their best interest. Those who understand **this**, will have an advantage.)

(d) a specifically marked link (*exoph* for exophor) denoting that the referent is "out" of the co-text, it is known only from the situation (ex. 8):

(8) *V období vrcholícího léta roku 1939 již málokdo v Evropě mohl uvěřit nadějeplným slovům Chamberlaina, pronesený* […] *po návratu z Mnichova: Myslím, že je **to** mír na celou naši dobu.*
(In the height of summer 1939 only a few people could believe the hopeful words Chamberlain uttered […] after the return from Munich. I think that **this** is peace for our time.)

(e) a specifical mark (*Unsp* for unspecified) is reserved for cases of reference difficult to be identified; this does not mean that a decision is to be made between two or more referents but that the reference cannot be specified even if the situation is taken into account (ex. 9):

(9) *Zmizení tohoto 700 kg těžkého přístroje* […] *hygienikům ohlásili (**Unsp**) 30. června letošního roku. Podle informací LN však zářič ze skladu Škody Plzeň zmizel již koncem letošního roku.*
(Lit.: The disappearance of the medical instrument weighing 700 kg […] **[they]** annonced on June 30th this year. According to the information of LN, however, the radiator […] disappeared by the end of the last year.)

1. It should be added that these cases occur only with "reconstructed" nodes in the TGTS's, i.e. with nodes that have no specific lexical value assigned, see ex. 9, in which the node for the Actor of the verb *oznámili* (annonce) is to be reconstructed but the antecedent is unspecified (who annonced? – whoever may be responsible for the annoncement)

2. The manual annotation is made user-friendly by a special tool in the TRED editor used for tree-structure assignment (see Kučová et al. 2003); the values of the attributes of coreference with each node of the tree will be assigned by an automatic procedure.

3.3 Until now, 717 PDT files of about 50 sentences each have been annotated as for the above types of textual coreference relations; the total number of sentences annotated is 34 272 and the total number of nodes (excluding the identification nodes for each of the sentences) is 429 155, out of which there are 14 658 anaphors of the type we have worked with (i.e. that are rendered by a personal or a demonstrative pronoun, possibly also a zero in the surface shape of the sentence, with the exclusion of the personal pronouns of the 1st and 2nd persons), see Table 1:

| | |
|---|---|
| number of annotated files | 717 |
| total number of sentences | 34 272 |
| total number of nodes (excl. the inentification node) | 429 155 |
| number of co-refering nodes (of the analyzed type) | 14 658 |
| % of co-refering nodes | 3, 4156 % |

**Table 1:** Volume of data

The distribution of the types of links (see above in Sect. 3.2) within the total number of 14 658 links is given in Table 2. The statistics demonstrates that a prevailing number of links has led to an explicit antecedent, while the number of exphoric relations is almost negligible. This might be due to the fact that most of the texts within the Czech National Corpus (from which the texts for the PDT collection were chosen) belong to the journalistic style, in which the reference to some explicit antecedent within the text itself is a standard stylistic strategy.

| | explicit antecedent | segm | exoph | unsp | total |
|---|---|---|---|---|---|
| number of links | 14 521 | 274 | 18 | 162 | 14 975 |
| % of the total | 96, 99 | 1,83 | 0,12 | 1,08 | 100 |

**Table 2:** Types of links

It may be also interesting to look at the distribution of the surface realization/deletion of the given type of anaphors: as Table 3 illustrates, the proportion of the expressed/restored anaphors is just 1 to 1; the number of personal pronoun lemmas (*on*) assigned to the anaphors (be they expressed or restored) is four times greater than the number of demonstrative pronoun lemmas (*ten*).

| | total | % |
|---|---|---|
| corefering nodes of the analyzed type | 14 658 | 100 |
| nodes expressed | 7 537 | 51,42 |
| nodes restored | 7 121 | 48,58 |
| lemma *on* (he) | 11 802 | 80,52 |
| lemma *ten* (that) | 2 856 | 19,48 |

**Table 3:** Some basic characteristics

In Tables 4 and 5 we present some statistics which we still plan to analyze in more detail because we hope to gain some interesting observations on the relation between coreference links and the underlying syntactic structure of the sentences; this may eventually help to formulate certain preferences for the selection of antecedents in an automatic procedure for the assignment of pronominal reference. The comparison of the values of functors (i.e. of underlying valency relations) with anaphors and antecedents indicates that the coreferential links hold mostly between inner participants (arguments) rather than between circumstantials (adjuncts): ACT(or), PAT(ient) and ADDR(essee) are among the three most frequent anaphors/antecedents. APP(urtenance) is a valency relation that typically belongs to the valency of nouns and as such is a relation of a dependent to its head noun, while the other relations in the Tables are those of dependents on verbs. The label PRED(icate) is assigned to the governing verb of the given TGTS and the figure in this column in Table 5 indicates that 6,11% of all coreferential links pointed to the governing verb of (one of) the preceding clause(s), which means that the antecedent is the event identified by the verb (be it together with some dependent nodes on this verb or not) of (one of) the preceding sentence. Neither of the Tables reflects from which functor to which functor the link goes, and this is exactly what we want to study further.

| | ACT | PAT | APP | ADDR | EFF |
|---|---|---|---|---|---|
| total | 8 092 | 3 103 | 1 276 | 568 | 326 |
| % | 55 % | 21,16 % | 8,07 % | 3,87 % | 2,22 % |

**Table 4:** Functors with anaphors

| | ACT | PAT | PRED | APP | ADDR |
|---|---|---|---|---|---|
| total | 6 839 | 3 015 | 916 | 864 | 627 |
| % | 45,67 % | 20,13 % | 6,11 % | 5,77 % | 4,19 % |

**Table 5:** Functors with antecedents

The total number of occurrences of the types of anaphoric links does not equal the total number of the occurrences of the anaphors, because there were cases in which a link has led to more than a single node; this situation can be illustrated by ex. 10, where the (superficially deleted) pronoun *oni* has as its antecedent both *tatínek* and *maminka*.[1]

[1] A technical remark: this treatment is necessary because in the construction *tatínek s maminkou šli*, *tatínek* stands in the relation of an Actor and *maminka* in the relation of Accompaniment to the verb *šli* rather than a coordination between two Actors; in case of true coordination, as in (11), and in case of apposition, the arrow leads to the node representing coordination (apposition) relation rather than to the members of the relation.

(10) *Tatínek s maminkou šli do divadla. Vzali **[oni]** si taxíka*.
(Father and mother went to the theatre. Took **[they]** a taxi.)

(11) *Tatínek, maminka a obě děti šli do divadla. Vzali **[oni]** si taxíka*.
(Father, mother and both children went to the theatre. Took **[they]** a taxi.)

## 4. Some interesting phenomena and open questions

The first phase of the coreference annotation process has revealed several interesting phenomena concerning anaphoric relations in Czech; in this Section we exemplify some problematic cases of textual coreference as present in real texts of PDT.

**4.1** The link labeled as *segm* covers also cases in which it is not quite clear where are the boundaries of the relevant segment or which concrete events/states in the previous segment are referred to, see ex. 12:

(12) *Jediný důvod k pobytu v Americe jsou pro mě peníze.* […] *Každý rok si v Americe najmu dům a po skončení sezony hned spěchám domů. Mám tu přátele, chodíme na ryby, hrát tenis, navštěvujeme se. Často jezdím za rodiči do Martina. Jsem tu prostě doma.* […] *V Kanadě je **to** úplně jiné.*
(The only reason for me to stay in America is money. […] In America, I rent a house every year and at the end of the season I rush home. I have friends here, we go fishing, we play tennis, we visit each other. I often visit my parents in Martin. I am simply at home here. […] In Canada **this** is totally different.)

Often it is not really relevant where the segment has its boundary, see (13): by what action the field has been prepared: by the minister's admission or by his opening the possibility?

(13) *Slovenský ministr kultury* […] *připustil, že zápůjčky obrazů nemusí být jednosměrné.* […] *Otevřel tedy možnost, o které se dosud nemluvilo. Ředitelům obou galerií **tím** zároveň připravil pole, na němž si mohou vzájemně ustoupit.*
(The Slovak minister of culture […] admitted that the loans of pictures need not be unidirectional. He thus opened a possibility which has not yet been discussed. By **this**, he prepared the field for the directors of both galeries so that they can make mutual concessions.)

**4.2** In complex cases exemplified here by (14), a decision should be made not only on the coreferential links but also on the restoration (and lexical labels, lemmas) of the nodes deleted in the surface shape of the sentence. The Czech verbs *říkat, sdělovat* (tell), *zapomenout* (forget), *ukázat* (show), *zapamatovat* (remember smth), *pochopit* (understand smth) have a semantically obligatory participant (argument) of Patient. This argument is deleted in the surface shape of (14) but it should be reconstructed in the respective TGTS. This reconstructed node receives the lemma *Gen* (general) in the first clause in each pair (except for the last pair where the demonstrative *to* (it) is present in the outer shape of the sentence), since it can be paraphrased as "everybody concerned" (with no coreferential link) and with the lemma *ono* (it) in the second clause of the respective pair because there its reference is not a general one but a link is to be established to the Patient of the first clause.

(14) *Každá kultura má svá rčení, která popisují zkušenosti lidstva s učením. České sděluje: Opakování, matka moudrosti. Čínské praví: Řekni mi **[Gen]** a já zapomenu **[ono]**; ukaž mi **[Gen]** a já si **[ono]** zapamatuji; nech mne **to** dělat a já **[ono]** pochopím.*
(Every culture has its own sayings, which describe the experience of mankind with learning. The Czech says: Repetition is the mother of wisdom. The Chinese say: Say **[it]** and I will forget **[it]**; show me **[it]** and I will remember **[it]**; let me do **it** and I will understand **[it]**.)

**4.3.** Along with clear cases of exophoric relations (exemplified here by (15)): one should know, at least from the history lessons at school, that the antecedent of the demonstrative pronoun is the Munich

Treaty) the corpus provides examples of **boundary case between exophora and other types of coreferential relations.**

(15) *V období vrcholícího léta roku 1939 již málokdo v Evropě mohl uvěřit nadějeplným slovům* […] *Chamberlaina, proneseným* […] *po návratu z Mnichova: Myslím, že je* ***to*** *mír na celou naši dobu.*
(In the height of summer 1939 only a few people could believe the hopeful words […] Chamberlain uttered […] after the return from Munich. I think that **this** is peace for our time.)

For instance, it is difficult to decide between the exophoric coreference, as e.g. in (16) and (17), a coreference to an unspecified element somehow deducible from the preceding context as e.g. in (18), or a co-reference to a segment (perhaps of the "inferential" kind, see ex. 19):

(16) *Na churáňovských svazích se* ***to*** *zelená, běžkaři na kvildských pláních masově krouží na posledních zbytcích vlhkého sněhu.*
(On the hills of Churáňov **[it]** looks green, the cross-country skiers on Kvilda plains make big circles on the last remains of wet snow.)

(17) *Děkuji za sérii povídání o Osvětimi. Jsem rád, že se konečně píše o tom, jak* ***to*** *skutečně bylo.*
(Thanks for the series of writings about Auschwitz. I am glad that finally one writes about how **it** really was.)

(18) *Největší tragédie se však stala v Pardubicích. Známý místní rodák Roman M., autor Průvodce pardubickými restauracemi, se upil k smrti po zjištění, že se narodil v Hradci Králové. Tento fakt vydedukoval z kopií žádostí svých rodičů, aby pardubická matrikářka zfalšovala Romanův rodný list. Rození pardubických dětí v Hradci Králové je periodicky se opakující jev. Jednou za dva roky nám je sem* ***[oni]*** *vozili, sdělila sestra na porodnickém oddělení hradecké fakultní nemocnice.*
(The worst tragedy was in Pardubice. A well-known native of Pardubice, Roman M., […] had drunk himself to death after he found out that he was born in Hradec Králové. He deduced this fact from a copy of the application of his parents. […] The birth of children from Pardubice in Hradec Králové periodically happens. Once in every two years **[they]** brought them here, said the nurse at the obstetric clinic of the Hradec hospital.)

(19) *Smutní lidé píší veselé knížky a veselí lidé smutné. V člověku se* ***to*** *musí nějak vyrovnat.*
(Sad people write bright merry books and merry people write sad [ones]. One has to balance **it** somehow.)

**4.4** A form of a demonstrative pronoun can be, of course, used in **other than referential functions**, as the following examples demonstrate:

(a) a demonstrative pronoun can be used as an **intensifying particle** *to* (with no coreferential link), see ex. 20:

(20) ***To*** *ale prší!*
(Boy, is it raining! Lit. **[that]** but it-rains! = meaning: it rains very much)[2]

(b) a **conceptually "empty"** occurrences of a form of the demonstrative pronoun (Šmilauer's "zdánlivý podmět/předmět" ("apparent" subject/object) is illustrated by (21) and (22):

(21) […] *jak si už dlouho představuju její cestu do ciziny, do Španělska nebo Řecka, kam ji* ***to*** *táhne.*

[2] Ex. 20 may be also used (with a different intonation!) in a context: „What's happening outside? It is raining.", in which *to* (it) is an exophor, refering out of the text.

([…] as I have imagined for a long time her trip abroad, to Spain or Greece, where [lit.] **it** draws her.)

(22) Všichni slzeli a radovali se tak z toho úspěchu, kam **to** dotáhl jeden z nich.
(All were crying and were glad of the success, where one of them [lit.] worked **it** [= worked oneself to])

(c) If a demonstrative pronoun is used in **phrasemes** or „frozen" collocations, no coreferential links are established; as a matter of fact, the form *to* (the neuter form for the demontrative *ten*) does not function as a pronoun here, see exx. (23) through (25):

(23) ***To** máte těžké, mladému to beztak obšlápnul táta.*
(Lit. **That** you-have hard, this young person's father has connections.)

(24) *Nevím, **čím to je**, ale absolutně se mi tady nedaří.*
(I do not know, **what's the matter**, but I am absolutely unsuccessful here.)

(25) *Mezitím do Pchanmundžonmu, odkud byli v dubnu vypuzeni pozorovatelé České republiky, přijíždí i mnoho Korejců a hledí nepřítomně do dálky, na sever. **Moc toho** ovšem v tomto prostoru **k vidění není**.*
(In the meantime, Pchanmundžonm, from where the Czech observers were expelled in April, is visited by many Koreans and they look absent-mindedly into the distance. There **is not much to see** in that area.)

**4.5** One of the advantages of a corpus-based study of a language phenomenon is the fact that the researchers become aware of subtleties and nuances that are not apparent. It is then desirable to collect a **list of open questions** which are handeld on the basis of a **temporary instruction** but which should be studied more intensively and to a greater detail in the future. The result, of course, is an **open list,** which is complemented during the whole course of the anntation process. The following examples illustrate what kind of problems we have encountered in our work:

(a) a coreferential link leads to the root of the tree but sometimes the antecedent is just a part of the whole sentence rather than the sentence (governed by the given verbal node) as a whole: in (26) the antecedent of *to* (this) is only the main clause *rozklepala se mi nejen kolena, ale i nitro* (not only my knees but also my heart trembled) rather than the whole complex sentence:

(25) *Když mi Jiří Krupička poslal rukopis své Renezance rozumu, která nyní vyšla v Českém spisovateli, a já do ní napoprvé nahlédl, rozklepala se mi nejen kolena, ale i nitro. A **to** hned z mnoha důvodů.*
(When Jiří Krupička sent me the manuscript of his Renaissance of Reason, which has been published now in the publishing house Český spisovatel, and I looked into it for the first time, not only my knees but also my heart trembled. And **this** [happened] for several reasons.)

(b) With a **coreferential chain**, all links (in the backward direction) are established, as in ex. 26; the link would lead from the last (superficially deleted) *on* (they) to the preceding (again superficially deleted) *on* (they), and from there to the preceding *on* (them) (expressed in the surface by the Acc. Pl. *je* and then finally to *protestanti* (protestants).

(26) *Dohoda pochopitelně nic nevyřešila – pouze prohloubila v **protestantech** pocit, že **je** Londýn nechává na holičkách. Dnes tento pocit, že jsou **[oni]** pro Británii pouze břemenem, s nímž si **[oni]** neví rady, v ulsterských protestantech pouze zesílil.*
(The agreement of course has not solved anything – it only deepened the feeling in the **protestants** that London leaves **them** in the lurch. Today this feeling, that **[they]** are only a burden for Great Britain, which **[they]** do not know how to deal with, has strengthened in Ulster protestants.)

(c) Since the determination of coreferential links is performed after the TGTS's have been built by other annotators, who are instructed not to take coreference into account when deciding on the lemmas with reconstructed nodes, it sometimes happens that the lemma assigned by these annotators is not an appropriate one from the point of view of coreferential relations. One option would be to **change** these **lemmas**, as indicated in (27) and (28):

(27) *Kluk odpočívá dlouze, nehnutě na jedné noze. Když přijdu blíž, postaví se na obě nohy a* ***řekne*** *mile: džambo, memsahib, how do you do?*
(The boy rests for a long time, very still on one leg. When I come closer, he stands on both legs and **says** pleasantly: jumbo, memsahib, how do you do?)

(28) *Tak báseň není pouze jen pytel slov, nespočívá toliko ve věcech, které znamená.* […] *A ryzí óda, jak tělo překrásné, sluncem i olejem zářící.* […] *Z výboru Múza* ***přeložil*** *Ivan Slavík.*
(So a poem is not just a sack of words, it is not anchored only in the things it denotes. […] A true ode as a beautiful body, shining with sun and oil. […] Ivan Slavík **translated** from the selection of poems Muza; Lit.: From […] translated Ivan Slavík.)

In (27) a (general participant) Addressee is restored (for the verb *řekne* [said]) with the lemma *Gen*, which would be during the coreference annotation changed to *on* because the antecedent has been found to be quite specific: *memsahib*. In a similar vein, in (28), a (general) Patient (for the verb *přeložil* [translated]) is restored with the lemma *Gen* and this lemma would be then changed to *on* (with a link to *segm*). Tentative criteria have been formulated guiding (and severely rerstricting) such changes: if it is possible to add some specific referent in place of the deleted node of the surface structure, this means that an arrow can be established pointing to some concrete node (or *segm*, as the case may be) obeying the general guidelines and *Gen* would be changed to o*n / ten*; else *Gen* would be left untouched.

(d) Nodes are reconstructed not only as dependent on a verb but also in cases of productively formed **nominalizations** if some of their obligatory complementation is deleted in the surface shape of the sentence; the establishment of coreferential links follows the same general guidlines, see (29) and (30):

(29) *[slovo] Má silné citové zabarvení a vyskytuje se zvláště v mluvených projevech mládeže.*
(It [= the word] has a strong emotive colouring and it occurs especially in discourse of young people.)

In the TGTS of (29), two nodes depending on *zabarvení* (colouring, from *zabarvit* [to colour]) are restored: both with the lexeme *Gen*, one with the functor Actor and one with Patient. In the course of the coreference annotation, the lemma *Gen* would be preserved with the Actor (there is no direct reference, meaning "anybody" colours…), *Gen*.Patient would be changed to *on* (with a link to *slovo* [word]).

(30) *Řekl jste, že občan ČR má po pěti letech od listopadu 1989 mnoho důvodů ke znepokojení, poukázal jste zvláště na vysoké daňové zatížení.*
(You said that five years after November 1989 a citizen of the Czech Republic has many reasons for disatisfaction, you pointed especially to a high tax load.)

In the TGTS of (30), again two nodes depending on *znepokojení* (dissatisfaction) are restored, namely *Gen*.Actor and *Gen*.Patient; the same happens with the restoration of two nodes with the deverbative *zatížení* (load). In the course of the coreference annotation, the lemma *Gen* would be preserved with the Actor of *znepokojení* (dissatisfaction) - there is no direct reference, meaning "anybody" dissatisfies - and *Gen*.Patient would be changed to *on* (with a link to to *občan* [citizen]). In the case of *zatížení* (load) both restored participants are left as "general", no referential link being established.

However, a more detailed analysis of these and similar cases is necessary to decide on the conditions under which a change of lemmas would be necessary. Therefore, in the present stage of annotation process, we have decided to keep the lemmas as they have been asssigned by the annotators of the syntactic structure untouched and to return to this issue in the future.

The annotation process has also revealed several other interesting phenomena concerning coreference in Czech, for example the issues of other than referential functions of pronouns (pronouns as intensifying particles) or a wide range of phrasemes and idioms. The study of these issues is open for further investigation.

## 5. Concluding Remarks

The approach to corpus annotation is a complex task performed in several levels and steps. The annotation of coreference relations is carried out on underlying (tectogrammatical) tree structures assigned to the sentences in the text on independent (and theoretically based) grounds, which makes it possible to systematically include into the annotation the superficially "null" (unrealized) anaphors and other phenomena not realized overtly in the surface shape of the sentences. The use of 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 34 272 sentences with the aim to assign the links and the values of the coreference attributes to the whole set of 50 000 sentences annotated on the underlying syntactic level). It should be emphasized that the coreference assignment as described here is not done selectively but it is an integral part of a large scale project of dependency-based annotation of underlying sentence structure (along with the annotation of the information structure of sentences) and as such it prepares solid grounds for further linguistic investigations.

## Acknowledgements

The research reported on in this paper has been supported by the project of the Czech Ministry of Education (MŠMT) LN00A063.

## References:


Berger Tilman. 1993. Das System der tschechischen Demonstrativpronomina. Habilitation. Ludwigs-Maxmilians-Universität, Munich.

Collins Michael. 1999. Head-Driven Statistical Models for Natural Language Parsing.

PhD Dissertation, University of Pennsylvania, Philadelphia.

Hajič Jan. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Vallency and Meaning. Studies in Honor of Jarmila Panevová*, ed. E. Hajičová, pages 106–132. Karolinum, Prague.

Hajič Jan, and Hladká Barbora. 1998. Tagging Inflective Languages: Predicting Morphological Categories for a Rich, Structured Tagset. In *Proceedings of the Coling '98*, pages 483–490. Montreal, Canada.

Hajič Jan, and Urešová Zdeňka. 2003. Linguistic Annotation: from Links to Cross-Layer Lexicons. In *Proceedings of the Treebanks and Linguistic Theories*, pages 69–80. Sweden.

Hajič Jan. 2004. Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Prague.

Hajičová, Eva. 2000. Dependency-based Underlying-structure Tagging of a Very Large Czech Corpus. In *Les grammairs de dépandance*, ed. Sylvain Kahane, pages 57–78. Paris.

Hajičová Eva, Panevová Jarmila and Sgall Petr. 2000. Coreference in Annotating a Large Corpus. In *Proceedings of LREC 2000*, volume 1, pages 497–500. Athens, Greece.

Hajičová Eva, Sgall Petr, and Veselá Kateřina. 2003. Information Structure and Contrastive Topic. In *Annual Workshop on Formal Approaches to Slavic Languages*. Ed. Brown, Wayles et al. Ann Arbor, Michigan Slavic Publications.

Kolářová Ivana. 2003. Zájmena v textu, jejich funkce deiktická a emocionální [Pronouns in text, their deictic and emotional function]. Poster at the conference "Čeština – univerzália a specifika". Brno.

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

Lopatková Markéta, Žabokrtský Zdeněk, Skwarska Karolina, and Benešová Václava. 2003. VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report of the Center for Computational Linguistics, Charles University, Prague.

Prague Dependency Treebank: http://ckl.mff.cuni.cz

Sgall Petr. 1979. Towards a Definition of Focus and Topic. The Prague Bulletin of Mathematical Linguistics 31, pages 3 – 25; 32, 1980, pages 24–32; printed in *The Prague Studies in Mathematical Linguistics*, 7, 1981, pages 173–198.

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

Šmilauer Vladimír. 1947. Novočeská skladba [Syntax of Modern Czech], Prague.

Štícha František. 1999. K deikticko-anaforickým funkcím lexému *ten* [On deictico-anaphoric functions of the lexeme ten]. In *Slovo a slovesnost* 60, pages 123–135.

TRED: http://ckl.mff.cuni.cz/pajas/tred