Acknowledgements

The work on this project was partially supported by the grants "Multilingual Corpus Annotation as a Support for Language Technologies" (LH14011, Ministry of Education, Youth and Sports) and “Coreference, discourse relations and information structure in a
contrastive perspective” (P406/12/0658, Czech science foundation).

Contents

Introduction

The Prague Czech-English Dependency Treebank is a parallel corpus manually annotated at the deep syntactic layer of linguistic representation. The English part consists of the Wall Street Journal (WSJ) section of the Penn Treebank. The Czech part was translated from the English source sentence by sentence. The detailed overview of the underlying linguistic theory (tectogrammatical annotation) with some details of the most important features like valency annotation, ellipsis reconstruction, etc. can be found in Hajič et al. (2012). In this report, we will present the annotation of coreference links in English (PEDT) and Czech parts of PCEDT.

Full annotation of textual coreference follows up the annotation of grammatical coreference in PEDT and completes the textual coreference taken for PEDT from the Ontonotes Release 4.0[1]. The rules and principles of this annotation are based on coreference annotation rules for Prague Discourse Treebank 1.0 (PDiT 1.0; Poláková et al. 2013) that are described in detail in the annotation manual *Annotation on the tectogrammatical level in the Prague Dependency Treebank* (Mikulová et al. 2005) for grammatical and pronominal textual coreference and in the special technical report Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank (Nedoluzhko et al. 2011) for nominal coreference and bridging relations.

In Section 1, we will present basic notions on coreference which will be used in the rest of the report. Section 2 will describe grammatical and textual coreference annotated in PCEDT. Special types of textual coreference will be also addressed in Section 2. Coreferring expressions that are subject to annotation are listed and described in Section 3. Referring and non-referring expressions are decribed in Section 4. Annotation principles and conventions are presented in Section 5. The annotation procedure and the annotation tool are described in Section 5. The relation between coreference annotation in PCEDT and coreference and bridging annotation in PDiT is also addressed in Section 6.

[1] Being sufficiently detailed, the OntoNotes textual coreference covers only cca. 1/5 of all PEDT data (BBN Technologies, 2006).

1. Basic notions

Two or more expressions are considered to be *coreferential* if they refer to the same discourse entity. In our annotation, equivalence of the head nouns is not a necessary precondition to call the expressions coreferential.

The expression to which a sentence element refers is called *antecedent*. The referring expression is called *anaphoric expression* or *anaphor*. Apart from these, the terms coreferring expression (element) and - coreferred expression (element) are also used. These terms are more general and disregard the position of the expressions in the text – as both the antecedent and postcedent can be coreferred expressions.

On the referential level, we speak about specifying and generic expressions. *Specifying*
expressions are those that are used to refer to a particular extra-linguistic entity. *Generic expressions* refer to types or prototypical objects.

2. Grammatical and textual coreference - definitions, examples and annotation rules in PCEDT

The coreference annotation in PCEDT 2.0 captures grammatical coreference, pronominal textual coreference and nominal (extended) textual coreference. The common property of *grammatical coreference* is that the relations appear as a consequence of language-dependent grammatical rules. Grammatical coreference comprises several subtypes of relations, which mainly differ in the nature of referring expressions (e.g. relative pronoun, reflexive pronoun, etc.). By *pronominal textual coreference* (where reference is not only expressed by grammatical means, but also via context), anaphor is expressed by personal, possessive or demonstrative pronouns or by textual ellipsis. *Nominal (extended) textual coreference* can be realised by repetitions, synonyms, paraphrasing, hyponyms/hyperonyms, etc. Unlike grammatical coreference, textual coreference often occurs between entities in different sentences. The distinction between grammatical and textual coreference is considered to be basic, and it is annotated separately in PCEDT.

Let's consider each type of coreference in more detail and exemplify them for English and Czech.

2.1 Grammatical coreference

There are two kinds of expression of grammatical coreference: either anaphor is expressed in the form of a pronoun (*Peter hurt himself*) or it is given by the syntactic structure of the sentence (*Peter_$_i$ wants [Ø_$_i$ to sleep]*.), thus being not expressed on the surface level but can be reconstructed on the tectogrammatical level.

The following types of grammatical coreference can be distinguished:

1. Coreference with reflexive pronouns. In this case anaphoric pronoun mostly refers to the closest subject, cf. the following example (1), where the reflexive pronoun *herself* corefers with the subject *daughter*, which corresponds to the Actor argument.

(1) *My daughter likes to dress herself without my help.*

2. Coreference with relative elements. Relative pronouns and pronominal adverbs introducing relative clauses are linked to their antecedent in the governing clause, cf. (2), where the relative expression *who* corefers with the noun *boy* modified by the dependent relative clause.

(2) *Alex is the boy who kissed Mary.*

3. Control (a type of grammatical coreference that arises with certain verbs, called control verbs, such as *begin, let, want*, etc.). The control relation arises, for example, with the elided subject of the infinitive *sleep* and the subject *Peter* in the sentence (3).

(3) *Peter_$_i$ wants [Ø_$_i$ to sleep]*.

This is such a coreferential relation between controller and controllee, that (i) the controller is a member of the valency frame of the governing verb, e.g. *Peter* is a

member of the valency frame of the verb *want*, (ii) the controllee (in our case the elided subject of the infinitive *sleep*) is a member of the valency frame of the infiitive dependent on the control verb (in our case the infinitive *sleep*) and (iii) the infinitive is a valency modification of the control verb, e.g. in the analyzed sentence, *sleep* is a valency modification of the verb *want*.

4. Coreference with verbal modifications that have dual dependency, e.g. (4).

(4) *Jan saw Mary$_i$ [Ø$_i$ run around the lake]*.

In this case, grammatical coreference concerns non-expressed arguments of verbal modifications with the so called dual dependency (e.g., passive participles, gerunds, infinitives). This is, for example, the case of coreference of unexpressed Actor of the infinitive *run* with the Patient *Mary* of the governing verb *saw* in (4).

2.1.1. Grammatical coreference in English and Czech

Grammatical coreference rules for English and Czech are very similar but not totally identical. Compare (5) with parallel grammatical coreference links in English and Czech and (6) - (7), where different syntactic structure of languages gives different results.

(5) *Fleet also noted that, unlike other banking companies in the Northeast, it has been only marginally hurt by <nonperforming loans> <that>* {coref_gram from <that> to <nonperforming loans>} *have resulted from the slumping regional real estate market*. = *Společnost Fleet zároveň uvedla, že na rozdíl od ostatních bankovních společností na severovýchodě byla jen okrajově zasažena <nesplacenými půjčkami>, <které>* {coref_gram from <které> to <nesplacenými půjčkami>} *byly následkem poklesu místního trhu s nemovitostmi.*

In (6), grammatical coreference is expressed by syntactic control in English, and by relative pronoun in Czech:

(6) *Based on the number of Mesa <shares> outstanding not already <#Cor.PAT>* {coref_gram to <shares>} *owned by StatesWest, the proposed takeover would have a value of about $15.3 million*. = *Na základě množství <akcií> společnosti Messa v oběhu, <které>* {coref_gram from <které> to <akcií>} *společnost StatesWest dosud nevlastní, bude mít navrhované převzetí hodnotu okolo 15.3 milionu dolarů.*

Not all relative expressions have corresponding relative translation in both languages, e.g., in (7) and Figures 1 and 2, Czech relative *což* has no appropriate equivalent in English, thus it is mostly not translated, or expressed by appositive construction.

(7) *Společnost Tucson Electric uzavřela v kompozitním obchodování na Newyorské burze cenných papírů na 20875 dolaru za akcii, <u>což</u>* {coref_gram to <dolar>} *je pokles o 25 centů*. = *Tucson Electric closed at $20.875 a share, down 25 cents, in New York Stock Exchange composite trading*.

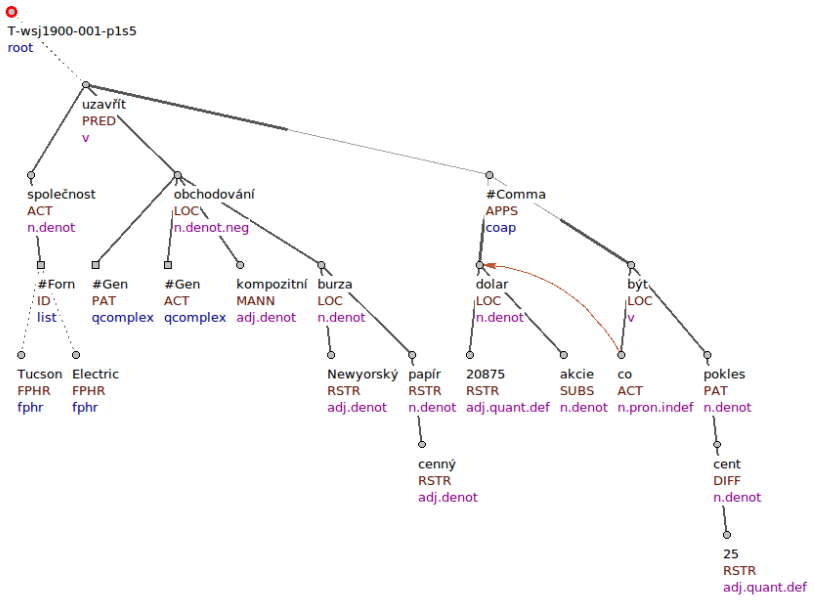

Fig.1. Grammatical coreference in English and Czech - cz

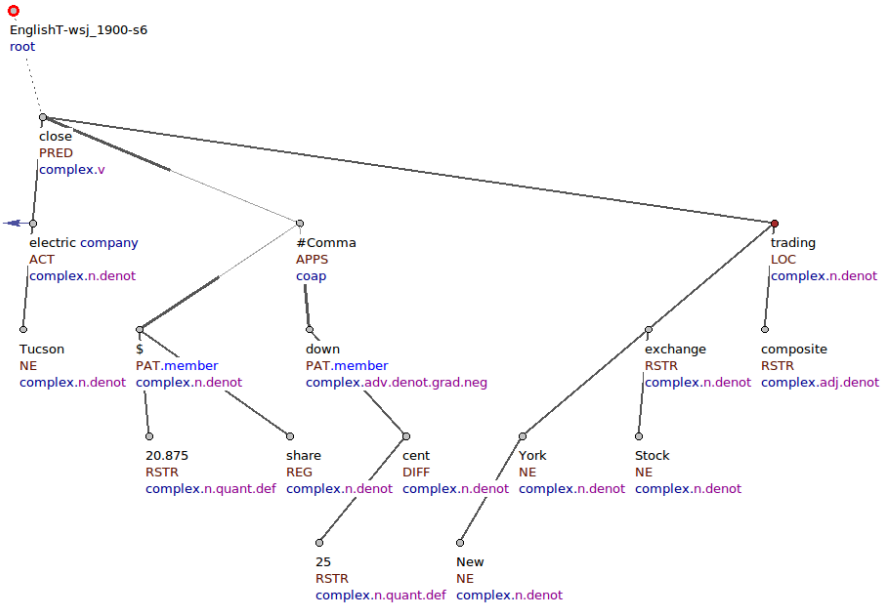

Fig. 2. *Grammatical coreference in English and Czech - en*

The comparison of coreferential expressions in English and Czech has begun in Novák - Nedoluzhko (2014) and continues to be a topic of further investigation.

2.2 Textual coreference

Textual coreference is annotated in the following cases:

1. Personal (in the 1st, 2nd, or 3rd person) and possessive pronouns. (In the tectogrammatical tree, personal and possessive pronouns have the single t-lemma #PersPron.)

> (8) *<A form of asbestos once used to make Kent cigarette filters> has caused a high percentage of cancer deaths among a group of workers exposed to <it> more than 30 years ago, researchers reported. = Výzkumníci uvedli, že <forma azbestu kdysi používaná k výrobě cigaretových filtrů značky Kent> způsobila vysoký podíl úmrtí na rakovinu mezi dělníky, kteří <jí> byli vystaveni před více než 30 lety*.

2. The demonstrative pronouns *this, that*.

(9) *They also said that <vendors were delivering goods more quickly in October than they had for each of the five previous months>. Economists consider <that> a sign that inflationary pressures are abating. = Ekonomové to pokládají za znamení polevujících inflačních tlaků. Uvedli rovněž, že <maloobchodníci dodávali v říjnu zboží rychleji než v každém z předchozích pěti měsíců>. Ekonomové <to> pokládají za znamení polevujících inflačních tlaků.*

3. With textual ellipsis, where a new node with the t-lemma substitute #PersPron is added to the tectogrammatical tree. In English, it is not so frequent as in Czech part.

(10) *<More common chrysotile fibers> are curly and are more easily <#PersPron> rejected by the body, Dr. Mossman explained. = Dr. Mossman vysvětlil, že obvyklejší chrysotilová vlákna jsou vlnitá a tělo je dokáže snáze vypudit.*

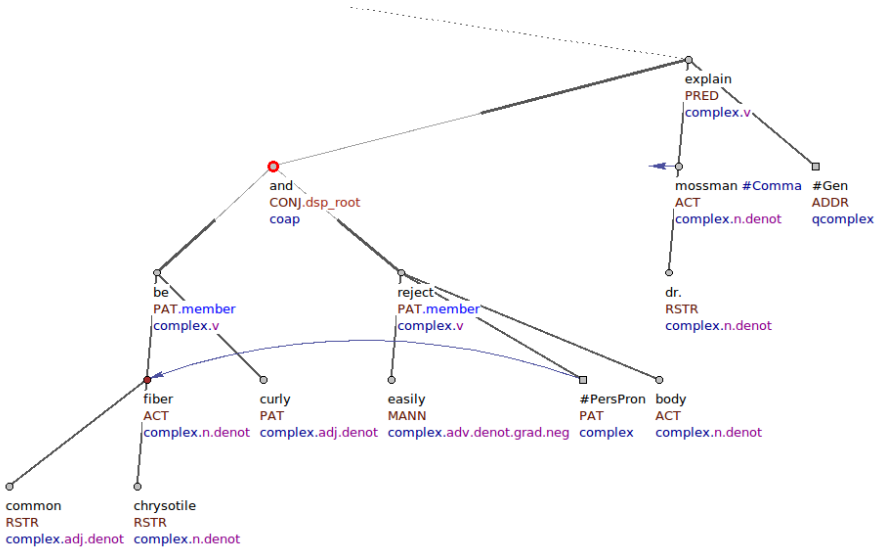

Fig. 3. Textual coreference, ellipsis

4. Nominal textual coreference. We do not annotate anaphoric relations in a restricted case, but we concentrate on marking the equivalence of referents of antecedent and anaphoric expressions. Cf., in (11), coreference is marked for the relation between *Fujitsu Ltd.* and *Japan's biggest computer maker*, although in English original text, the noun phrase *Japan's biggest computer maker* contains no

explicit anaphoric reference to the antecedent *Fujitsu Ltd.* It is interesting, however, that in the Czech translation, the anaphoric reference (*Tento největší počítačový výrobce v Japonsku* = lit. *This Japan's biggest computer maker*) is used.

> (11) *Japanese companies have long been accused of sacrificing profit to boost sales. But <Fujitsu Ltd.> has taken that practice to a new extreme. <Japan's biggest computer maker> last week undercut seven competitors to win a contract to design a mapping system for the city of Hiroshima's waterworks. = Japonské společnosti jsou již dlouho obviňovány z toho, že se vzdávají zisku, aby zvýšily obrat. Ale <firma Fujitsu Ltd.> tuto praxi dovedla do nového extrému. <Tento největší počítačový výrobce v Japonsku> minulý týden nabídl nejnižší cenu v porovnání se svými sedmi konkurenty a získal kontrakt na projekt mapovacího systému pro zásobování města Hirošimy vodou*.

Textual coreference is marked up to the length of 20 sentences. Annotating coreference for a greater number of sentences is possible only in cases of automatic pre-annotation of named entities coreference. This decision was made in order to avoid a large number of mistakes and to reach higher inter-annotator agreement.

!!! We did NOT annotate textual coreference in case of relations between wh-words and replies to them (e.g. *When will you call me? - In the evening*.)

!!! We do not distinguish between coreference pairs with the same lemmas (*Mary - Mary*) from the cases, in which the entities are synonymous, hyponymous / hyperonymous or are just different nominations of any other kind (*Germany – the state, Mary – she*, etc.). Using grammatical attributes of the tectogrammatical tree, this kind of information can be easily extracted automatically. We also do not annotate false positive links (lexically identical but non-coreferential NPs) as coreferential.

Extended textual coreference may take place between different types of specifying NPs: the same t-lemmas (as in the example 12), different t-lemmas (example 13),

different t-lemmas, a kind of hyperonymous relation, different subtrees with the same governing node, and so on.

(12) *State court Judge Richard Curry ordered <Edison> to make average refunds of about $45 to $50 each to <Edison> customers who have received electric service since April 1986, including about two million customers who have moved during that period. = Soudce státního soudu Richard Curry nařídil společnosti Edison vrátit v průměru zhruba 45 až 50 dolarů každému zákazníkovi, který využíval její služby dodávky elektřiny od dubna 1986, včetně asi dvou milionů zákazníků, kteří se během tohoto období přestěhovali.*

(13) *The Illinois Supreme Court ordered the commission to audit <Commonwealth Edison's> construction expenses and refund any unreasonable expenses. <The utility> has been collecting for the plant's construction cost from its 3.1 million customers subject to a refund since 1986. = Illinoiský nejvyšší soud nařídil komisi, aby provedla audit stavebních výdajů <společnosti Commonwealth Edison> a vrátila veškeré nepřiměřené výdaje. <Tento podnik> veřejných služeb vybírá poplatky na výstavbu elektrárny od 3,1 milionu svých zákazníků s nárokem na refundaci od roku 1986.*

2.3 Special types of textual coreference (coref_special)

Three special types of relations are annotated in PCEDT together with the annotation of textual coreference:

- references to discourse entities external to the text (coref_special, type exoph), see 2.3.1;
- references to a discourse segment consisting of more than one sentence (coref_special, type segm), see 2.3.2; and
- coreferential relations with more than one antecedent, see 2.3.3.

2.3.1 References to discourse entities external to the text (Exophora)

In exophora, an expression refers to situations or reality external to the text. We are aware of the fact that the term coreference is usually used only for endophoric reference; still the annotation of exophoras is technically included into the coreference annotation.

Exophoric reference is represented by the attribute coref_special, which contains the value `exoph`. In the extended textual coreference annotation, only definite noun phrases or pronouns may be annotated for exophora.

Exophoric reference is annotated in the following cases:

- Time and local deixis, e.g. (14) and (15):

  (14) *Preliminary tallies by the Trade and Industry Ministry showed another trade deficit in October, the fifth monthly setback <this year>, casting a cloud on South Korea's export-oriented economy. = Předběžné záznamy Ministerstva průmyslu a obchodu ukázaly v říjnu další deficit obchodní bilance, v letošním roce již pátý měsíční pokles, což na jihokorejskou ekonomiku orientovanou na export vrhá stín.*

  (15) *The irony is that the attack commercial, after getting a boost in last year's presidential campaign, has come of age in an off-off election year with only a few contests scattered across <the country>. = Ironií je, že útočná volební reklama, která byla podpořena v loňské prezidentské kampani, dozrála ve volebně chudém roce, kdy se volby v USA pořádají jen na velmi málo místech.*

- Deixis with pronominal adverbs, e.g.

  (16) *<Here> are the Commerce Department's figures for construction spending in billions of dollars at seasonally adjusted annual rates. = <Zde> jsou údaje ministerstva obchodu o stavebních výdajích v miliardách dolarů v sezónně upravených ročních mírách.*

- Exophoric reference to the whole text, e.g.
  (17) *Informace <v tomto přehledu > jsou bezplatnou službou podnikatelům.*
  =The information <in this report> is a free service to businessmen.

Exophoric reference is annotated only in case of actual deixis (one can imagine that the
speaker is pointing with the finger by saying the phrase). For that reason, exophora is NOT annotated in the following cases:

1. Exophoric meaning is part of lexical semantics of a given expression (*dnes (= today),*
zítra (= tomorrow), současnost (= the present) etc.)
2. In syntactic constructions with deictic semantics, e.g. *příští rok (= next year), v současné době (= nowadays), minulý týden (= last week), v sobotu (= on Saturday), v červenci (= in June)* etc.
3. By reference to generic “we”, e.g.
   (18) *Zákon o prostituci se <u nás> teprve připravuje. (=* lit. *A law on prostitution is still being prepared at ours [meaning, in our country])*
4. By exophoric references to characteristics, e.g.
   (19) *Angel říká, že fronty se každým dnem znatelně prodlužují. „Viděl jsem šňupat opravdový dámy, jsou tu i lidi, který vypadaj, jako by umírali na AIDS. Je hrozný, jak jim takovýhle život {no coreference relation} užírá rozumný myšlení rychleji než blesk.“ (=Angel says that the queues are getting significantly longer. "... there are also people who look as if they were dying on AIDS. It is terrible to see how such a life eats their reasonable thinking even faster than lightning. ")* (example from PDiT)

2.3.2. References to a discourse segment consisting of more than one sentence

Reference to a segment takes place in the following cases:
• a noun phrase refers to a substantial section of a text consisting of more than one sentence (see 2.3.2.1),

• a noun phrase refers to a tree segment which cannot be technically separated (see 2.3.2.2)

Reference to a segment is represented by the attribute coref_special, in which the value `segm` is marked.

Reference to a segment does not have an explicit antecedent. It is supposed to be supplied in the future versions of coreference annotation.

2.3.2.1 Reference to more than one sentence (discourse deixis)

The cases of discourse deixis, where the anaphoric expression refers to one sentence, a clause or a verbal phrase, are described in Section 2.2. Here, we concern only reference to more than one sentence.

One speaks of reference to a segment in cases where a noun phrase (often with a determiner) refers to more than to one sentence in the previous context.

> *(20) In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed. About 160 workers at a factory that made paper for the Kent filters were exposed to asbestos in the 1950s. Areas of the factory were particularly dusty where the crocidolite was used. Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used to make filters. Workers described ``clouds of blue dust'' that hung over parts of the factory, even though exhaust fans ventilated the area. ``There's no question that some of those workers and managers contracted asbestos-related diseases,'' said Darrell Phillips, vice president of human resources for Hollingsworth&Vose. ``But you have to recognize that <these events> took place 35 years ago. = V červenci zavedla Agentura na ochranu životního prostředí postupný zákaz prakticky všech využití azbestu. Do roku 1997 budou postavena mimo zákon téměř všechna zbývající užití karcinogenního azbestu.Azbestu bylo vystaveno v padesátých letech v továrně vyrábějící papír pro cigarety Kent asi 160 dělníků. Oblasti továrny, kde se používal*

krokydolit, byly obzvlášť zaprášené. Dělníci při postupu používaném k výrobě filtrů vysypali velké jutové pytle dovezeného materiálu do velkého zásobníku, přidali bavlněná a octová vlákna a tato suchá vlákna mechanicky promíchali.

Dělníci popisovali "mračna modrého prachu", která se vznášela nad částmi továrny, ačkoliv odsávací větráky oblast provětrávaly. "Není pochyb, že se někteří z těchto dělníků a manažerů nakazili nemocemi spojenými s azbestem," řekl Darrell Phillips, viceprezident společnosti Hollingsworth & Vose pro lidské zdroje. "Ale musíte uznat, že se <tyto události> odehrály před 35 lety.

2.3.2.2 Reference to a tree segment which cannot be technically separated

There are some rare cases, where there is no technical possibility to separate the antecedent sub-tree. For the time being, such cases are also annotated as coref_special, type = segm, cf. example (21) and Figure 4.

(21) *Od 1. dubna nebude ÚNMS SR rozhodnutí české zkušebny potvrzovat. <Tato funkce> přejde na příslušnou slovenskou zkušebnu, která bude vydávat na základě dodaných podkladů příslušné certifikáty*. (= *From 1 April, the ÚNMS SR will not make confirmations to the decisions of the Czech department. b. <This function> will come to to the relevant Slovak rehearsal...*) (example from PDiT)

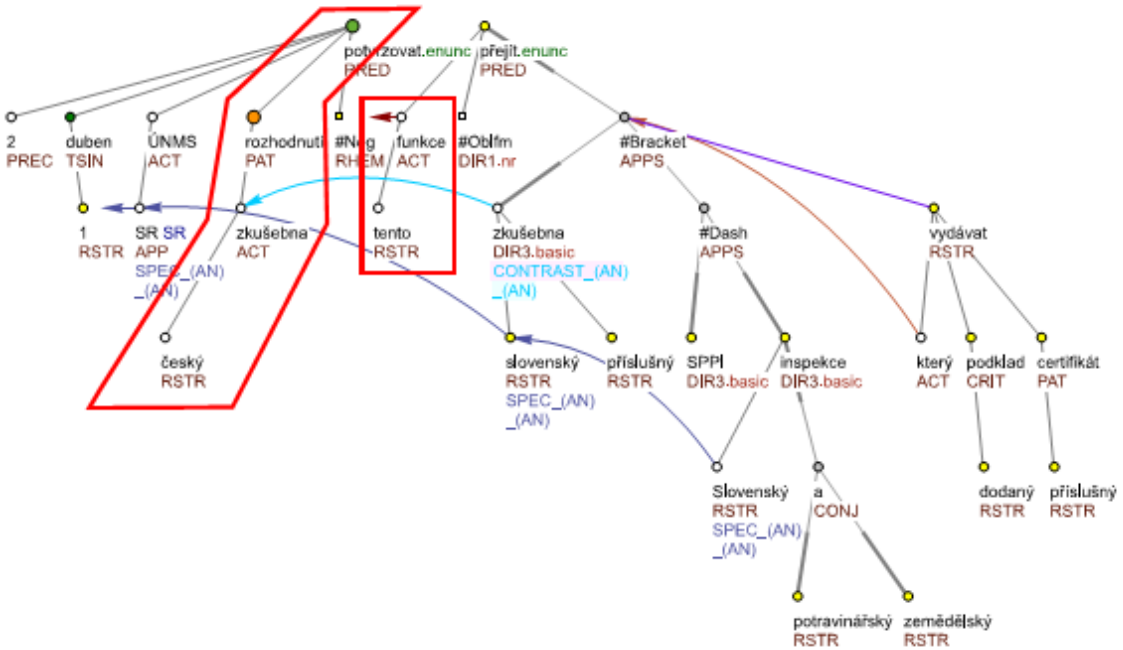

Fig. 4: Reference to a tree segment which cannot be technically separated

2.3.2.3 Two (or more) nodes of the tectogrammatical tree are antecedents of the anaphoric element.

This is the case of so called “split antecedent”, which is in PCEDT resolved as special type of relation. If C is coreferential with the sum of antecedents A+B, both present in tectogrammatical structure of the corresponding text, this relation is annotated. Such reference is represented by the attribute `bridging`, in which the value SUB_SET is entered.

For example, anaphoric NP *the companies* in the third sentence in (22) refers to two antecedents, <*Cray Research*> and <*Cray Computer*>.

> (22) *Under terms of the spinoff, <Cray Research> stockholders are to receive one <Cray Computer> share for every two Cray Research shares they own in a distribution expected * to occur in about two weeks. No price for the new shares has been set. Instead, <the companies> will leave it up to the marketplace to decide. = Podle podmínek odtržení mají akcionáři společnosti <Cray Research> získat jednu akcii společnosti <Cray Computer> vždy za dvě akcie společnosti Cray Research, které vlastní, a to při rozdělování, jež se očekává asi do dvou týdnů. Pro nové akcie nebyla dosud stanovena žádná cena. Namísto toho ponechají <společnosti> na trhu, ať rozhodne.*

3. Elements to be annotated

In this section, we describe elements which are subject to annotation for coreferential relations in PCEDT. Our classification is based on the part-of-speech classification, using the terminology used for annotation of tectogrammatical level in Mikulová et al. (2005), and the ability of elements to refer.

By classifying coreferential pairs, we look at the formal characteristics of the anaphoric expression. Considering the coreferential relation to be symmetric, the same is true for the formal characteristics of the antecedent. The exception is the coreferential relation with a situation (expressed by a verbal phrase). It has a different semantic interpretation than common coreferential relation, and thus it cannot be considered to be symmetric.

Coreferential relations are to be marked between elements of the following categories:

• Complex nodes in the anaphoric position - nodes representing autosemantical lexical units, pronouns, ellipses, etc. (see 3.1),

• Paratactic structure root nodes in the anaphoric position (3.2),

3.1 Complex nodes in the anaphoric position

According to Mikulová et al. (2005), we distinguish four basic groups of semantic word classes which are further subdivided – semantic nouns, semantic adjectives, semantic adverbs and semantic verbs. Semantic parts of speech are categories of the tectogrammatical level and correspond to the basic onomasiological categories: substances, properties, circumstances and events. Information about semantic parts of speech of a complex node is included in the attribute `sempos`.

3.1.1 Semantic nouns in the anaphoric position

Semantic nouns are the most frequent subjects for coreference annotation. The following groups of nouns are annotated:

1. Pronouns and demonstratives

Pronouns and demonstratives are linked to their antecedents, including pronouns in quoted speech.

> (23) *<Dobiaš> skoro všechno dělá s námi, <jeho> pověstná impulzivnost se přenáší i na nás, a to je dobře. = <Dobiaš> does almost everything with us; <his> notorious spontaneity carries over to us as well, and that is a good thing.*

Expletive pronouns (*it, there*) and generic ‘you’ are not linked. In the following example (24), the pronoun ‘you’ would not be marked:

> (24) *Senate majority leader Bill Frist likes to tell a story from his days as a pioneering heart surgeon back in Tennessee. A lot of times, Frist recalls, <you>’d have a critical patient lying there waiting for a new heart, and <you>’d*

want to cut, but <you> couldn't start unless <you> knew that the replacement heart would make it to the operating room.

2. Specific nominal mentions

Specific noun phrases are subject to annotation. Cf. the coreference chain for different nominations of *Aluminum Co. of America* in (25).

(25) *<Aluminum Co. of America, hit hard by the strength of the dollar overseas>, said net income for the third quarter dropped 3.2% to $219 million. <The nation's No. 1 aluminum maker> earned $226.3 million, or $2.56 a share, a year earlier. [...] Analysts, who were expecting <Alcoa> to post around $2.70 to $3 a share, were surprised at the lackluster third-quarter results. [...] Lower prices for aluminum ingots and certain alloy products and a shift in the product mix also contributed to lower earnings, <the company> said. = <Firma Aluminum Co. of America>, tvrdě zasažena silnou pozicí dolaru v zámoří, uvedla, že čistý příjem za třetí čtvrtletí se snížil o 3,2 % na 219 milionů dolarů. <Největší národní výrobce hliníku> vydělal o rok dříve 226,3 milionu dolarů, neboli 2,56 dolaru na akcii. Analytici, kteří očekávali, že <Alcoa> vynese přibližně 2,70 až 3 dolary na akcii, byli neradostnými výsledky za třetí čtvrtletí překvapeni. <Firma> řekla, že nižším výnosům napomohla i nižší cena hliníkových ingotů, jisté produkty ze slitiny a posun ve směsi produktů*.

3. Generic nominal mentions

Generic nominal mentions are linked to referring pronouns and other definite mentions, but not to generic nominal mentions. This instruction agrees with the rules of coreference annotation in Ontonotes release 4.0 and it allows linking of the bracketed mentions in (26) and (27), but not (28).

(26) *<Officials> said <they> are tired of making the same statements*. (OntoNotes)

(27) *<Meetings> are most productive when <they> are held in the morning. <Those meetings>, however, generally have the worst attendance*. (OntoNotes)

(28) *He said that <Jews> have contributed more to black causes over the years than vice versa. [...] He said <Jews> were ``sick with complexes''; and he called David Dinkins, Mr. Giuliani's black opponent, ``a fancy shvartze with a mustache." = Řekl, že <Židé> přispěli během let k černošským případům více, než tomu bylo naopak. Řekl, že <Židé> jsou "nemocní komplexy" a Davida Dinkinse, černého oponenta pana Giulianiho, nazval "líbivým negrem s knírkem."*

In example (29) below, there are three generic instances of ‘parents’. These are marked as three distinct coreferential chains, each containing a generic and the referring pronouns.

(29) *<Parents>x should be involved with their children's education at home, not in school. <They>x should see to it that <their>x kids don't play truant; <they>x should make certain that the children spend enough time doing homework; <they>x should scrutinize the report card. <Parents>y are too likely to blame schools for the educational limitations of <their>y children. If <parents>z are dissatisfied with a school, <they>z should have the option of switching to another*.

The same rule applies to indefinite nominal mentions in anaphoric position. In (30), the verb cannot be linked to ‘a reduction of 50%’, since ‘a reduction’ is indefinite.

(30) *Argentina said it will ask creditor banks to [halve] its foreign debt of $64 billion -- the third-highest in the developing world . Argentina aspires to reach <a reduction of 50%> in the value of its external debt*.

However, it is not so seldom that indefinite noun phrases refer to rather definite discourse entities, indefinite article being caused by other, for example, stylistic reasons. Cf. reference of *a Soviet bank* in (31):

(31) Coincident with the talks, the State Department said it has permitted <a Soviet bank> to open a New York branch. The branch of the Bank for Foreign

Economic Affairs was approved last spring and opened in July. But <a Soviet bank> here would be crippled unless Moscow found a way to settle the $188 million debt, which was lent to the country's short-lived democratic Kerensky government before the Communists seized power in 1917. = … pobočku v New Yorku. Tato pobočka <Banky> pro zahraniční ekonomické záležitosti byla schválena minulé jaro a otevřena v červenci. Avšak <zdejší sovětská banka> by byla ochromena, kdyby Moskva nenašla způsob, jak vyrovnat dluh ve výši 188 milionů dolarů, které si vypůjčila krátce trvající Kerenského demokratická vláda předtím, než se v roce 1917 chopili moci komunisté.

4. Premodifiers

Premodifiers that are common names are not supposed to be subject of coreference annotation in Ontonotes, and this rule was also borrowed for PCEDT. However, the large-scale annotation has shown that human annotators whose mother tongue lacks grammatical category of definiteness, tend to mark coreference in evident cases. Cf. the relation between *The Arizona Corporations Commission* and *commission* in *commission hearing* in (32). In Czech, this discourse entity is syntactically represented as common noun, thus being normally annotated for coreference.

(32) *<The Arizona Corporations Commission> authorized an 11.5% rate increase at Tucson Electric Power Co., substantially lower than recommended last month by a <commission> hearing officer and barely half the rise sought * by the utility. The Arizona regulatory ruling calls for $42 million in added revenue yearly, compared with a $57 million boost proposed by the commission hearing officer. = <Arizonská komise pro korporace> schválila společnosti Tucson Electric Power Co. zvýšení sazby o 11,5 %, což je podstatně méně, než minulý měsíc doporučoval jednající úředník komise, a sotva polovina zvýšení požadovaného podnikem. [...] Arizonské regulační nařízení požaduje roční zvýšení příjmů o 42 miliónů dolarů, přičemž jednající úředník <komise> navrhoval pozvednutí o 57 miliónů dolarů*.

Premodifiers that are proper nouns are linkable, unless they are in a morphologically adjectival form, e.g. (33).

(33) *<Hiroshima city> officials couldn't be reached to find out whether they would drop Fujitsu's bid. [...] Fujitsu said it hopes the Hiroshima contract will help it secure pacts with other municipalities. [...] No foreign companies bid on the <Hiroshima> project, according to the bureau. [...] Three competitors bid between 300,000 yen and 500,000 yen, according to the Hiroshima government office.*

Adjectival forms of geographical names such as 'Chinese' in 'the Chinese leader' are not annotated for coreference. Thus, in 'the <United States> policy' the proper noun is linked to other references, but not in 'the American policy'. Nationality acronyms are considered adjectival as well; i.e., *U.S.S.R.* or *U.S.*[2].

(34) *But <the Army Corps of Engineers> expects the river level to continue falling this month (...) the flow of the Missouri River is slowed, an <Army Corps> spokesman said. Acronymic premodifiers are co-referenced unless they refer to nationality*.

In the expression *<FBI> spokesman*, *FBI* can be coreferenced to other mentions, but 'U.S.' in *<U.S.> spokesman* cannot.

Even when acronymic nationality-premods act like their non-acronymic counterparts, they are not considered proper premodifiers. In (35), 'Japan' can be coreferenced, but 'U.S.' cannot[3]:

(35) *On U.S.-<Japan> relations: "I'm encouraged."*

Proper pre-modifiers that include acronyms in the span, however, are eligible for coreference:

(36) *A <U.S. Treasury> spokesman*

[2] See BBN Technologies (2006), s.4.
[3] See BBN Technologies (2006), s.4.

Pre-modifying dates and monetary amount are also coreferenced[4]:

(37) *The current account deficit on France's balance of payments narrowed to 1.48 billion French francs ($236.8 million) in August from a revised 2.1 billion francs in <July>, the Finance Ministry said. Previously, the <July> figure was estimated at a deficit of 613 million francs.* (Ontonotes)

(38) *The company's <$150> offer was unexpected. The firm balked at <the price>*. (Ontonotes)

5. Nested Proper Names

Proper names are considered to be atomic, and nested mentions inside proper names are not annotated separately, unless they are proper mentions themselves. In the following examples, the location names that form part of the organization names are not eligible for coreference[5].

6. Coreference with named entities

If coreferring expression is a named entity and coreferred expression is a common noun, which has the named entity as a direct dependent node with the ID or NE functor, coreferential relation is annotated to/from the governing node of the common noun, NE and ID functors are ignored. If a named entity consists of several words and refers to one object, coreference arrow marks the governing node. By dependent nodes, coreference is not marked. Cf. the following example and Fig. 5.

(39) *Coincident with the talks, the State Department said it has permitted a Soviet bank to open a New York branch. The branch of the Bank for Foreign Economic Affairs was approved last spring and opened in July. - Současně s rozhovory oznámilo americké ministerstvo zahraničních věcí, že povolilo sovětské bance otevřít pobočku v New Yorku. Tato pobočka Banky pro zahraniční ekonomické záležitosti byla schválena minulé jaro a otevřena v červenci*.

[4] See BBN Technologies (2006), s.5.

[5] This is a change with respect to BBN Technologies (2006). There, the proper mentions *Massachusetts*, *America* and *Chicago* are not annotated inside the proper names *Massachusetts Institute of Technology*, *Bank of America* or t*he Chicago Board of Trade*.

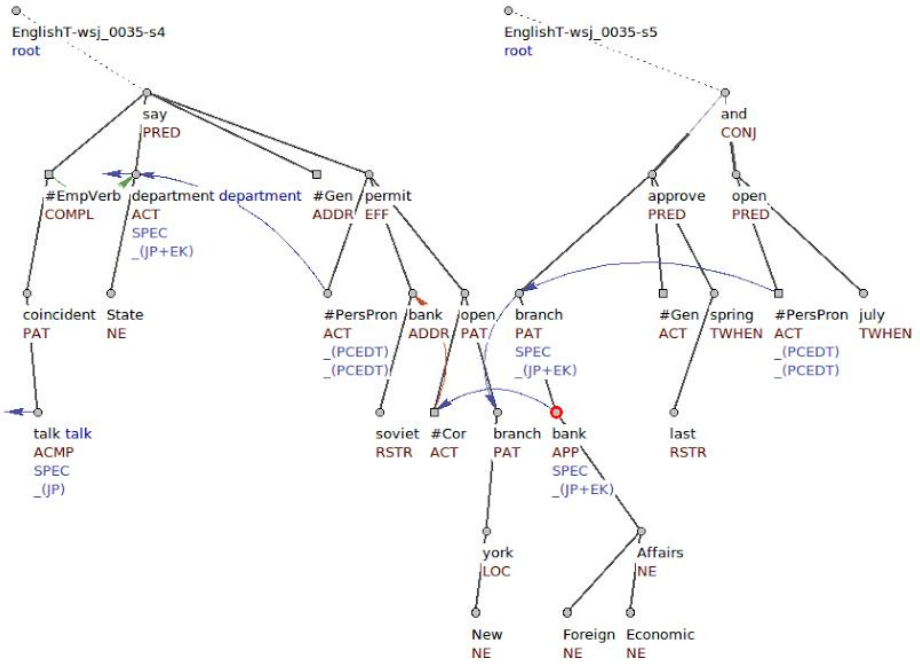

Fig.5 Coreference with named entities.

3.1.2 Semantic adjectives in the anaphoric position

Adjectives are not subject to annotation of coreference. We do NOT annotate:

1. Proper adjectival premodifiers[6] (*American, Australian*).

2. Proper abbreviations of place names, such as *U.S.* in the example below, are also not marked in premodifier position, since they are adjectival in nature. (*U.S. economy = American economy)*

3. Demonstratives in the position of semantic adjectives (*We didn't like <u>that</u> house*),

5. Indefinite pronouns in the position of semantic adjectives (*<u>Which</u> book did he want to have?*),

6. Numerals in the position of semantic adjectives (*He bought <u>five</u> books*).

7. Denominating semantic adjectives (*red, personal*, etc.).

3.1.3 Semantic adverbs in the anaphoric position

Out of all groups of semantic adverbs (see Mikulová et al.. 2005), only definite pronominal semantic adverbs (such as *there, here, then*) are annotated for coreference in PCEDT.

[6] For more detail, see 3.1.1

3.1.4 Semantic verbs as members of a coreferential relation

Verbs are not annotated for coreference as anaphors. Yet, semantic verbs (verbal phrases, clauses, sentences with a verb in the root, the whole situation described by more than one sentence) may still be antecedents of noun phrases in the anaphoric position. In this case, they are annotated as antecedents of coreferential relations.

(40) *Jistotu v tomto směru dávají nejnovější kroky vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience. b. Byť má <na tento krok> {coref_text to „zavést“} určité právo, v daném okamžiku však vyznívá jako tvrdé politické rozhodnutí vlády, která se snaží velice rezolutními administrativními kroky zredukovat mnohamilionové pasívum v obchodní výměně s ČR. (= <In this respect>, confidence can be derived from the newest steps of the Slovak government, which decided to introduce the previously announced 10% tax on goods imported from abroad. b. Even though it has the right to make this step {coref_text to „introduce“}, at this stage...)* (PDiT)

(41) *Sales of passenger cars grew 22%. The strong growth {coref_text to “grow”} followed year-to-year increases* (Ontonotes).

(42) *Japan's domestic sales of cars, trucks and buses in October rose 18% from a year earlier to 500,004 units, a record for the month, the Japan Automobile Dealers' Association said. <The strong growth> {coref_text to “rise”} followed year-to-year increases of 21% in August and 12% in September 7 .* (Ontonotes)

3.2. Paratactic structure root nodes in the anaphoric position

Root nodes of paratactic structures may be conjunctions used with coordination and opposition, e.g., *and, but*, t-lemma substitutes for syntactically relevant punctuation marks (e.g.: #Comma, #Dash, #Colon, #Separ, see Mikulová et al. 2005) or symbols

referring to mathematical operations and intervals. Paratactic structure root nodes are common as coreferring and coreferred elements.

When choosing the antecedent by annotating coreference in sentences with coordination and apposition structures, annotation to the whole structure, i.e. technically to a paratactic structure root node, is preferred. For example, in (43) and Fig. (6), the personal pronoun <*it*> refers to the whole appositional structure <*The big semiconductor and computer maker*>.

(43) <*The big semiconductor and computer maker*>, *said* <*it*> *had net of [...]. = Tento velký výrobce polovodičových součástek a počítačů uvedl, že dosáhl čistého zisku [...].*

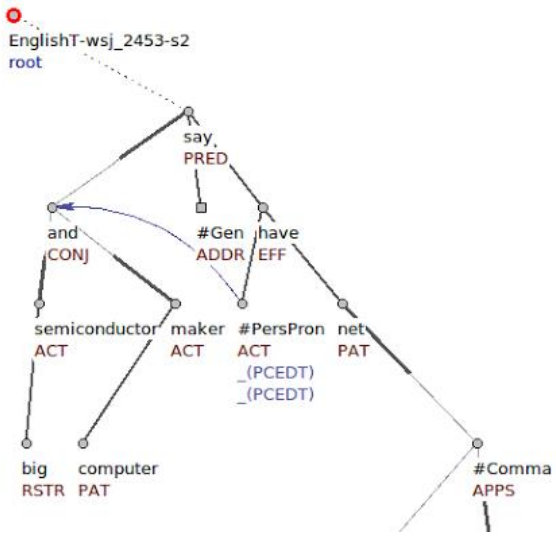

Fig. 6. Paratactic structure root nodes in the anaphoric position

4. Referring and non-referring noun phrases

In the annotation of textual coreference, we distinguish between referring and non-referring noun phrases. Non-referring NPs are not to be annotated.

The following NPs are considered to be non-referring:

1. Predicative NPs, except for identification constructions, where the predicative part of the sentence may be the antecedent for the anaphoric phrase in what follows. So, e.g. the relation between *Petr* and *programmer* in the sentence *Petr is a programmer*

is NOT annotated as coreference. In the same way, coreference is not marked in identification structures *Petr is the carpenter who did our bathroom*. This decision has been made, because this relation is already included in the tectorgammatical structure and can be easily extracted if needed.

2. Noun phrases, which form the second parts of appositions (e.g. no coreference relation between *Andrew S.Grove* and *Intel president and chief executive officer* in (44) and Fig. 7.)

(44) *On Friday, Andrew S.Grove, Intel president and chief executive officer, said ``Intel's business is strong. = V pátek řekl Andrew S.Grove, prezident Intelu a výkonný ředitel, že "Obchod Intelu je silný*.

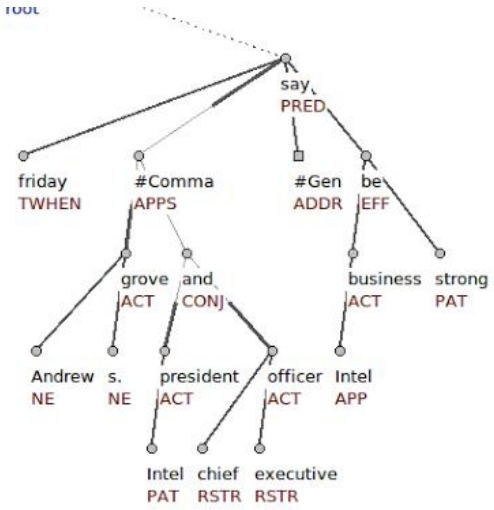

Fig. 7. Non-referring noun phrases

3. Identifying expressions, which are represented as identification structures (in the

tectogrammatical structure, they have the functor ID or NE).

4. Other non-referring NPs, such as measures, points etc. in contexts like the following:

(45) *Ve stejném období minulého roku činil čistý příjem společnosti SCI 4.8 <milionu> dolarů, neboli 23 centů za akcii, při příjmech ve výši 225.6*

<milionu> dolarů. = In the year-earlier period, SCI had net income of $4.8 <million>, or 23 cents a share, on revenue of $225.6 <million>.

5. Annotation principles and conventions

In order to develop a maximally consistent annotation scheme, we follow a number of basic principles. Some of them are presented below:

Chain principle

Coreference relations in text are organized in ordered chains. The most recent mention of an entity is marked as the antecedent. This principle is checked automatically.

The principle of maximum size of anaphoric expressions

By annotating coreference relations, the principle of maximum size of an anaphoric expression was applied. It is always the whole subtree of the antecedent/anaphor which is subject to annotation. Technically, coreference arrows go from/to the governing nodes of the coreferring expressions. Cf. in (46), the whole expression, i.e. *Aluminum Co. of America, hit hard by the strength of the dollar overseas* is subject to coreference annotation. See Figure 8.

(46) *<Aluminum Co. of America, hit hard by the strength of the dollar overseas>, said net income for the third quarter dropped 3.2% to $219 million. = <Firma Aluminum Co. of America, tvrdě zasažena silnou pozicí dolaru v zámoří>, uvedla, že čistý příjem za třetí čtvrtletí se snížil o 3,2 % na 219 milionů dolarů, neboli o 2,46 dolaru na akcii*.

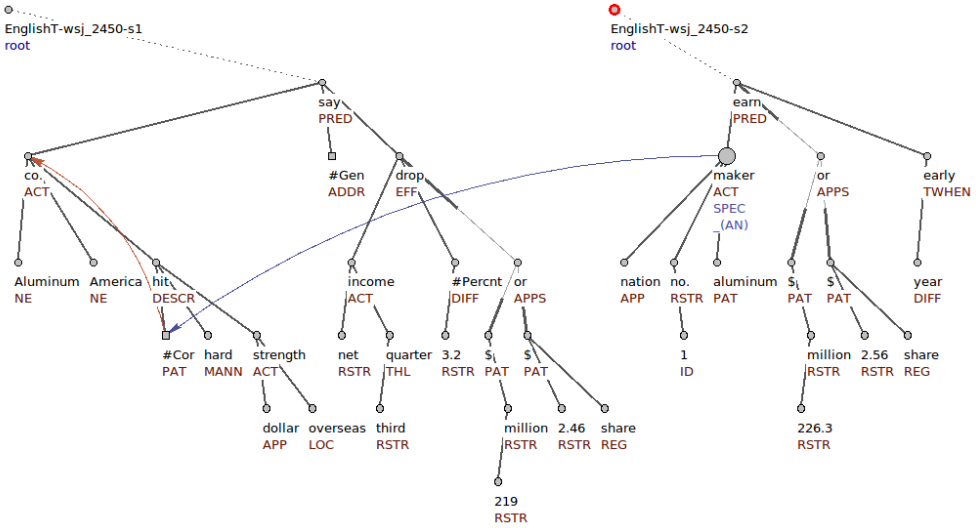

Fig.8. The annotation principles

Annotation of textual coreference is based on the chain principle, the anaphoric entity always referring to the last preceding coreferential antecedent.

Principle of maximal size of an anaphoric expression.

This principle says that it is always the whole subtree of the antecedent/anaphor which is the subject to the annotation.

Principle of cooperation with the syntactic structure of the given dependency tree.

We do not annotate relations that are already captured by the syntactic structure of the tectogrammatical tree.

Principle of preferring coreference to anaphora.

Coreference, not anaphora, is subject to textual coreference annotation. In many cases, an anaphoric relation is also a coreferential relation, this is however not always the case.

6. Realization of coreference annotation in PCEDT

For the English part of PCEDT (for PEDT), the resulting manual pronominal coreference annotation was built above an automatic transformation of the original coreference annotation extracted from BBN Pronoun Coreference and Entity Type Corpus (LDC2005T33). It has been further manually checked and corrected. Nominal coreference has been extracted from BBN Pronoun Coreference and Entity Type Corpus for the part for which it has been completed. The rest (about 80% of the PEDT has been annotated manually).

For the Czech part of PCEDT, textual coreference has been manually annotated, independently from the English texts.

6.1 Relation of coreference annotation in PCEDT to coreference and bridging annotation in PDiT

Coreference annotation in the Czech part of PCEDT is completed according to the rules of coreference annotation in PDiT (Nedoluzhko 2011), however simplified as follows:

- Only noun phrases with specific reference have been annotated for coreference. Hence, there have been no type specification. Opposite to this, in PDiT, coreference has been annotated including generics, with further specification to coreference for specific and generic noun phrases (Nedoluzhko, 2011)
- In PDiT, bridging relations have been considered. In PCEDT, bridging relations were not included, except for one special case, described in 2.3.2.3.

6.2. Annotation in TrEd

The annotation format of PCEDT 2.0 is called PML. It is an abstract XML-based format designed for annotation of treebanks. For editing and processing data in PML format, a fully customizable tree editor TrEd has been implemented (Pajas & Štěpánek 2008).

TrEd can be easily customized to a desired purpose by extensions that are included

into the system as modules. In this section, we describe some features of an extension that has been implemented for our purposes.

The data scheme used in PCEDT has been enriched to support the annotation of textual coreference. Technically, various kinds of non-dependency relations between nodes in PCEDT use dedicated referring attributes that contain unique identifiers of the nodes they refer to.

Visualisation

The following Figure 9 shows the basic features of the coreference annotation. Coreference relations between subtrees are marked by arrows of different colors (dark-red arrows for grammatical coreference and dark-blue arrows for textual coreference), the arrow pointing from an anaphor to an antecedent. If an antecedent is found in one of the preceding sentences, its lemma is written in dark-blue next to its anaphor.

(47) *He spends his days sketching passers-by, or trying to. Tráví den kreslením portrétů kolemjdoucích či se o to alespoň snaží*.

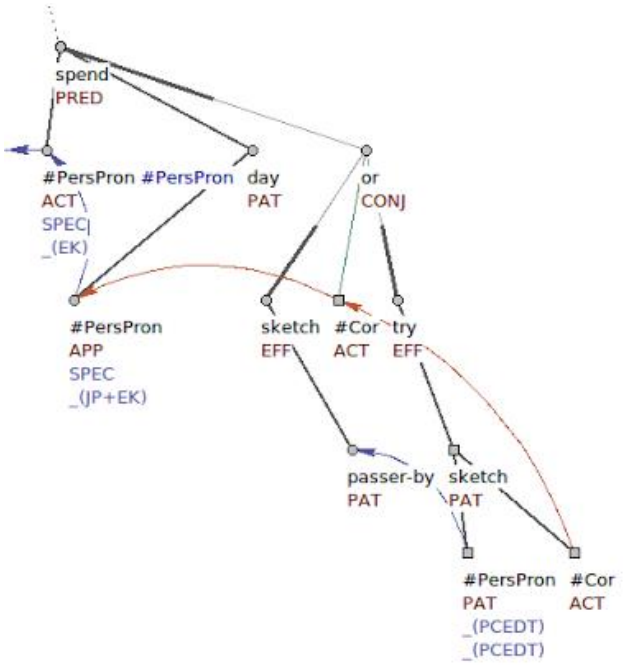

Fig.9. Visualisation

The Annotation

Several features have been implemented in the annotation tool to help with the annotation during the annotation process.

Manual pre-annotation

If annotators find a word in the text that appears many times in the document and its occurrences seem to co-refer, they can create a coreferential chain out of these words by a single key-stroke. All nodes that have the same t_lemma become a part of the chain.

Finding the nearest antecedent

The annotation instructions require that the nearest antecedent is always selected for the coreferential link. The tool automatically re-directs a newly created coreferential arrow to the nearest one (in the already existing coreferential chain) if the annotator selects a farther antecedent by mistake. However, the rule of the nearest antecedent can be broken in less clear situations. For example, if there are three coreferential words in the text, A, B and C (ordered from left to right), and the annotator connects A and C (overlooking B), and later realizes that B is also coreferential with A and creates the arrow from A to B, the tool re-connects the C→A arrow to C→B. Thus, the coreferential chain C→B→A is correctly created. Cf. the following Fig. 10:

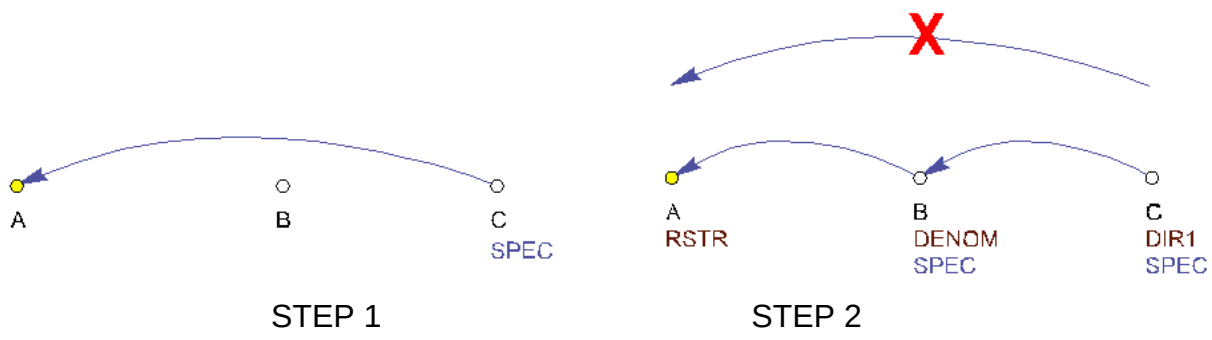

Fig. 10: Finding the nearest antecedent

Preserving the coreferential chain

If the annotator removes an arrow and a coreferential chain is thus interrupted, the tool asks the annotator whether it should re-connect the chain, as shown at the Fig. 11:

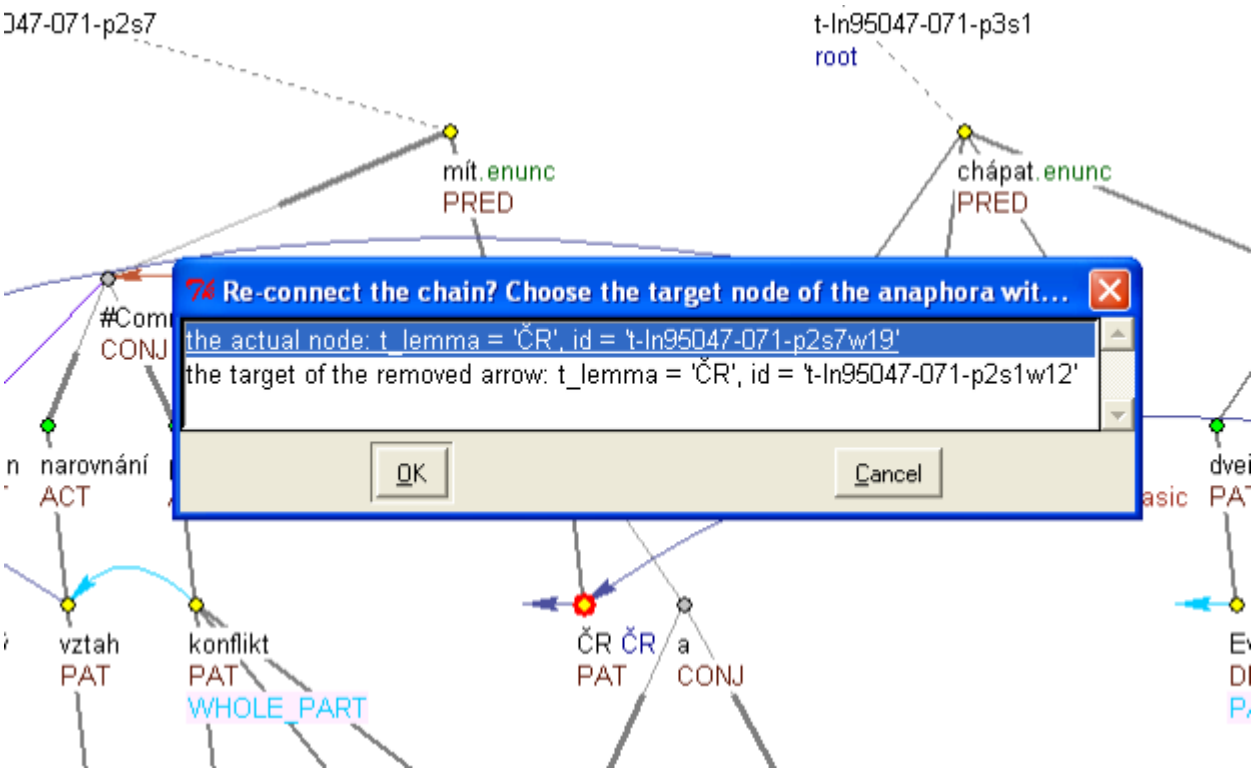

Fig. 11: Preserving the coreferential chain

Text highlighting

The annotation of coreference is performed on the tectogrammatical layer of PCEDT. However, the annotators may work on the surface form of the text too, using the tectogrammatical trees only as a supporting depiction of the relations. After selecting a word in the sentences (by clicking on it), the tool determines to which node in the tectogrammatical trees the word belongs. Then, the projection back to the surface is performed and all words on the surface that belong to the selected node are highlighted. Only one word of the highlighted words is a lexical counterpart of the tectogrammatical node (which is usually the word the annotator clicked on – only in cases such as if the annotator clicks on a preposition or other auxiliary word, the lexical counterpart of the corresponding tectogrammatical node differs from the word clicked on). Using this information, also all words in the sentences that have

the same t_lemma (again, we use only the lexical counterparts) as the selected word, are underlined. Words that are connected with the selected word via a coreferential chain are highlighted in such colors that indicate whether the last connecting relation in the coreferential chain was textual or grammatical. Moreover, all words that are connected via a bridging anaphora with any word of this coreferential chain, are highlighted in a specific color. Cf. Fig. 12:

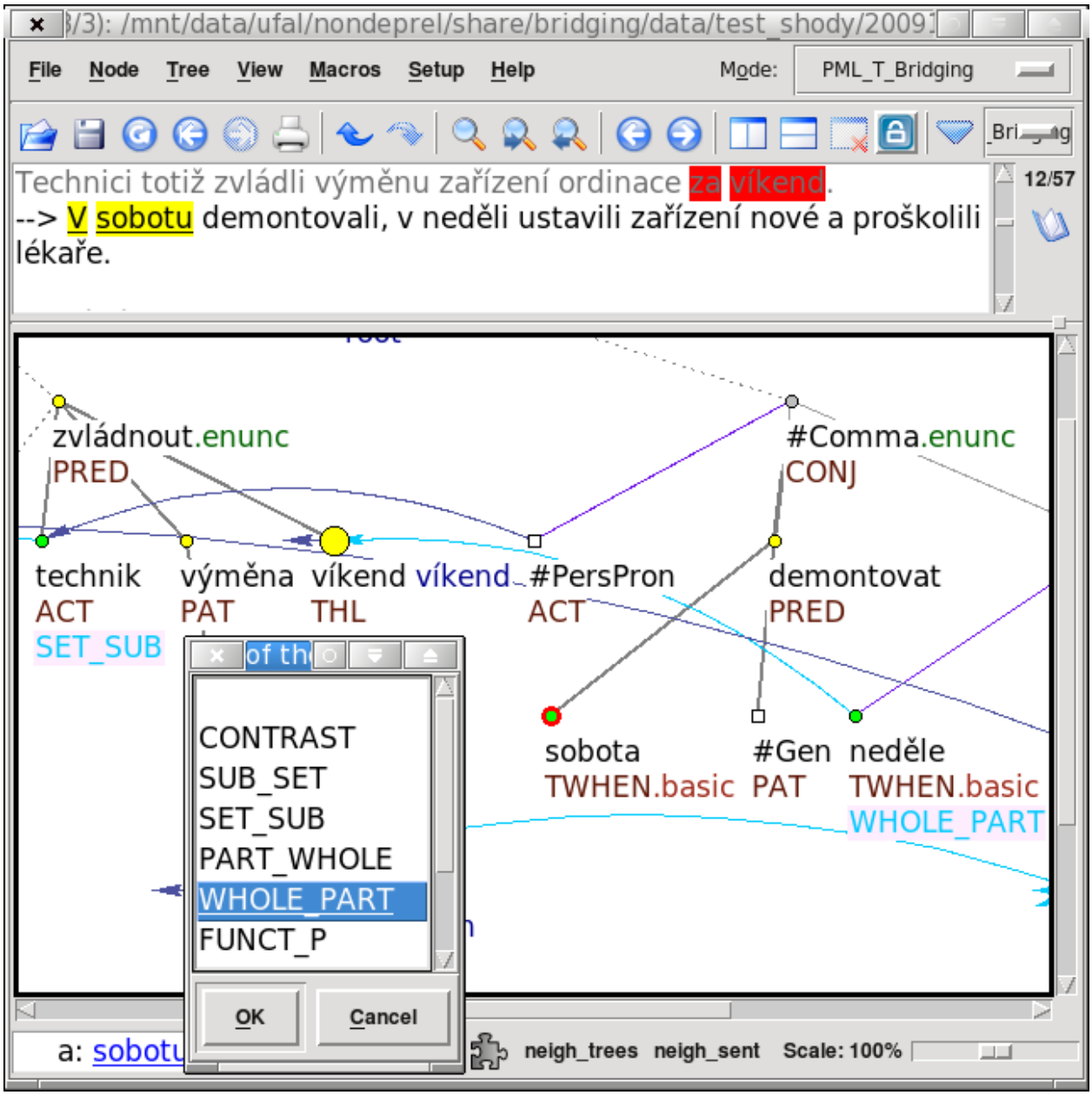

Fig. 12: Text highlighting

Comparing different annotations

The tool provides a support for visual comparison of different annotations of the same data, e.g. annotations from different annotators in the inter-coder agreement measurement.

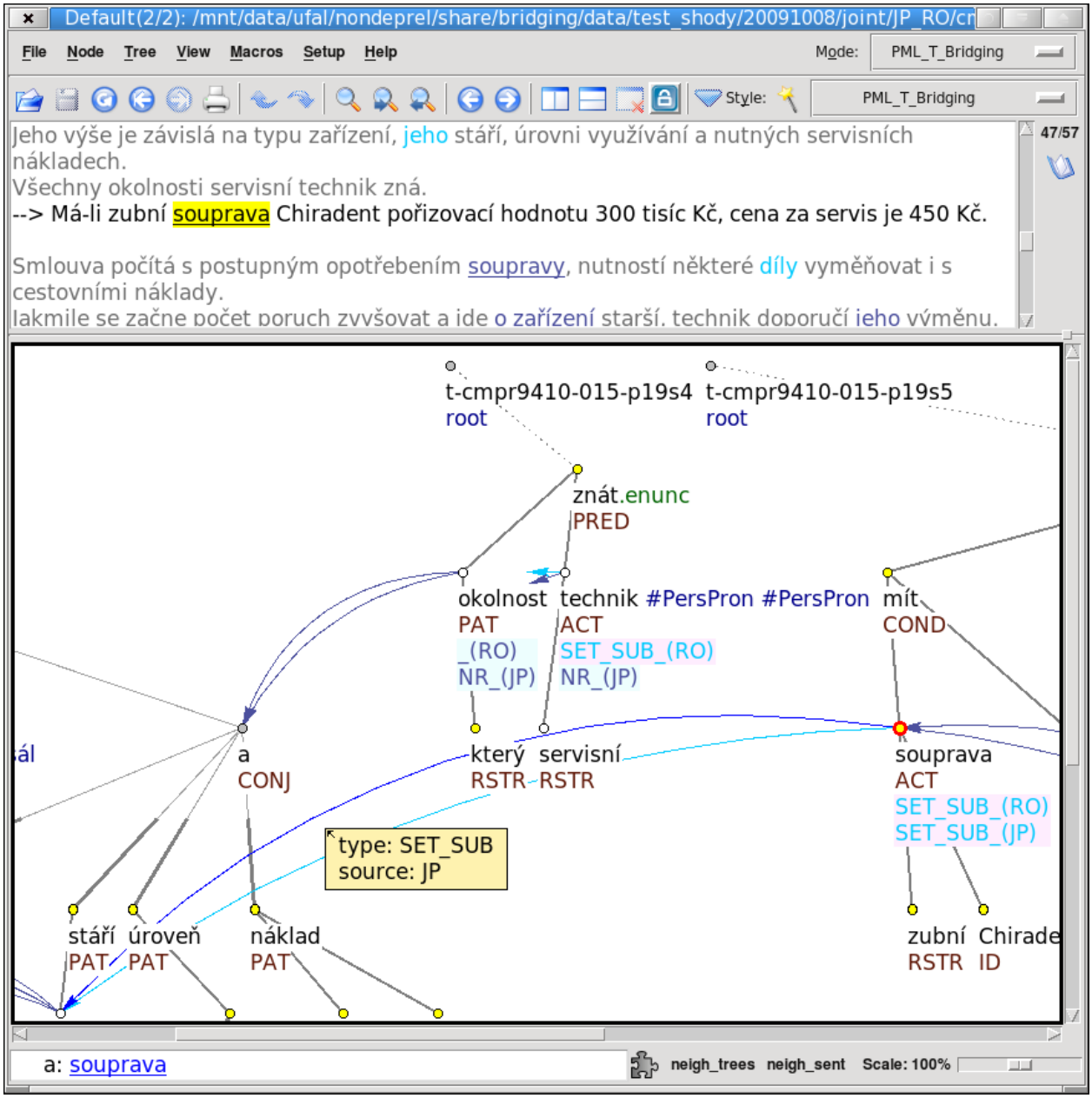

Fig. 13: Comparing different annotations

References

BBN Technologies. Coreference Guidelines for English OntoNotes – Version 6.0. Linguistic Data Consortium. BBN Pronoun Coreference and Entity Type Corpus. 2006. (LDC2005T33)

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová and Zdeněk Žabokrtský. *Announcing Prague Czech-English Dependency Treebank 2.0*. LREC. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey

Mikulová, Marie et al. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka, I, II. Technická zpráva ÚFAL TR-2005-28. Praha: Universitas Carolina Pragensis, 2005.

Nedoluzhko, Anna; Mírovský, Jiří. 2011a. *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank.* Annotation manual. Technical report No. 44, ÚFAL, Charles University in Prague, 2011, 63 pp.

Novák, Michal, Nedoluzhko, Anna. Comparison of coreferential expressions in Czech and English. In *Discourse*, 2014 (*to appear*)

Pajas, Petr; Štěpánek, Jan. Recent advances in a feature-rich framework for treebank annotation. In Proceedings of the The 22nd Interntional Conference on Computational Linguistics. Manchester, 2008, s. 673–680.

Poláková, Lucie, Mírovský, Jiří, Nedoluzhko, Anna, Jínová, Pavlína, Zikánová, Šárka, Hajičová, Eva. 2013. Introducing the Prague Discourse Treebank 1.0. In: Proceedings of the 6th International Joint Conference on Natural Language Processing, Copyright © Asian Federation of Natural Language Processing, ISBN 978-4-9907348-0-0, pp. 91-99.

Poláková, Lucie, Jínová, Pavlína, Zikánová, Šárka, Hajičová, Eva, Mírovský, Jiří,

Nedoluzhko, Anna, Rysová, Magdaléna, Pavlíková, Veronika, Zdeňková, Jana, Pergler, Jiří, Ocelák, Radek. 2012. Prague Discourse Treebank 1.0. Data/software, ÚFAL MFF UK, Prague, Czech Republic, Nov 2012 (downloadable CD distribution)