Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank

Anna Nedoluzhko, Jiří Mírovský

This technical report describes the project of manual annotation of extended textual coreference and bridging relations, which runs at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, since 2009. It contains the typology of coreference and bridging relations, classification of elements that are annotated for coreference and the application of the annotation on PDT 2.0.

This work has been supported by the GAČR 405/09/0729 grant.

Table of Contents

1 Introduction

The present technical report describes the annotation of Czech data for extended nominal coreference and bridging relations in the Prague Dependency Treebank, version 2.0 (PDT 2.0). This annotation is a continuation of the annotation of grammatical and textual coreference that was completed for PDT 2.0 in 2003. The annotation of extended nominal coreference and bridging relations is based on the tectogrammatical level (the annotation has been carried out on on tectogrammatical trees and many elements of tectogrammatical level, such as functors, node types, grammatemes etc. have been made use of), so the basic knowledge of tectogrammatical terminology is needed to understand the text.

In Section 2, the basic notions used in the manual are described. The annotation of the previous stage of grammatical coreference (4.1) and pronominal textual coreference (4.2) is briefly presented with examples. In Section 3, some tectogrammatical functors relevant to the annotation of coreferential and bridging relations are listed. Annotation principles and preferences are presented in Section 6. The annotation scheme for extended textual coreference and bridging relations is presented in Sections 7 and 9, respectively. Special cases of textual coreference appear in Section 8. Problematic cases that could not be included in the previous chapters are analyzed in Section 10. Technical representation of the annotation is described in Section 11.

2 Basic notions

Two or more expressions are considered to be *coreferential* if they refer to the same extra-linguistic entity. In our annotation project, 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*.

Coreferential relations in PDT are mainly anaphoric, but there are some rare cases, where a coreferential arrow leads cataphorically to the subsequent expressions; such expression can be called *postcedent.* (e.g. *Tu nejvhodnější dobu pan Hrabák propásl. V osmdesátých letech se daly pořídit krásné věci za, viděno dneškem, ještě krásnější ceny*. (*= Mr. Hrabák passed on the (*lit. *that) most opportune moment. In the eighties, one could get the most beautiful things for, from today's perspective, even more beautiful prices.*)).

Apart from these, other terms are also used: *coreferring expression* (*element*) - *coreferred expression (element)*. 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.

Non-coreferential association relations are called *bridging* if they stand in some semantic, lexical or conceptional relation to their antecedent.

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.

3 Tectogrammatical functors relevant for annotation of coreference and bridging relations

3.1 The PREC functor

The PREC functor refers to preceding text; it represents an expression linking the clause to the preceding context.

Expressions with the PREC functor are mainly subject to annotation of discourse relations (Mladová – Zikánová – Hajičová 2008); they are not expected to refer, and therefore the expressions with the PREC functor are not annotated for coreference and bridging relations.

E.g.

> *a. Jsem si jist, že ne hony na čarodějnice, ale právě poukázání a následné právní kroky, učiněné vůči opravdovým pachatelům násilí a komunistické svévole, je povzbudivým signálem toho, že naše společnost sice pozdě, ale přece jen vyvodila praktický a konkrétní krok, potvrzující přesvědčení, že historii tvoří konkrétní lidé, kteří ve svých činech projevují svou svobodnou vůli a nesou tedy za ně i svou osobní odpovědnost. b. Ukazuje se <u>tak</u>* {no coreferential relation}*, že pokud společnost chce, může najít onu pomyslnou dělicí čáru. (= I am certain that not the witch hunts, but actually the identification of and subsequent legal action against the real perpetrators and communist free-for-all, is an encouraging signal of the fact that our society, although late, has made practical steps that confirm the belief that history is made by specific people who show their free will through their actions and, as such, carry responsibility for them. <u>This</u> {no*

coreferential relation} shows that if society wants to, it can find the potential dividing line.)

3.2 The ID functor

The functor ID (identity) is used as a functor for an identifying expression, which is represented as an identification structure. The ID functor is assigned to adnominal adjuncts representing meta-language expressions, proper nouns and names of animals, objects and events, e.g. *v případu Kott - Kutílek (= in the case of Kott-Kutílek)*; *agentura Reuters (=Reuters agency)*; *pojem čas (=notion of time)*. In such cases, noun phrases do not refer to objects but to themselves (so-called autonymous reference in the terminology of Padučeva (1985)).

3.3 The ACMP functor

The ACMP functor (accompaniment) is a functor for such an adjunct which expresses manner by specifying a circumstance (an object, person, event) that accompanies (or fails to accompany) the event or entity modified by the adjunct.

The meaning of the ACMP functor may come in conflict with some bridging relations (mainly SUBSET, see 9.1.2). In this case, the bridging relations are not annotated. Cf. *válečná plavidla včetně bojových letadel.ACMP a bitevních vrtulníků.ACMP (=warships including air force ...)*

4 Coreference already annotated in PDT 2.0

In the available version of PDT 2.0, the following types of coreference are fully annotated:

- Grammatical coreference (4.1);
- (Pronominal) textual coreference (4.2).

Annotation of extended textual coreference and bridging relations follows up the annotation of grammatical and pronominal textual coreference, completed for PDT 2.0 in 2003. The rules and principles of this annotation are described in detail in the annotation manual *Annotation on the tectogrammatical level in the Prague Dependency Treebank* (Mikulová et al. 2005) and in the special technical report *Anotování koreference v Pražském závislostním korpusu* (Kučová et

al. 2003). In the present paper, we briefly characterise grammatical coreference in 4.1 and pronominal coreference in 4.2.

4.1 Grammatical coreference

Grammatical coreference is a kind of coreference in which it is possible to identify the antecedent on the basis of grammatical rules. In the case of grammatical coreference, both antecedent and anaphor are located in the same sentence.

The following types of grammatical coreference can be distinguished:

1. Coreference with reflexive pronouns. Cf. the following example where the reflexive pronoun *sobě* corefers with the subject *matka* (=*mother*), which corresponds to the Actor argument.

   Sobě nedopřeje matka nikdy nic. (=lit. To_herself not_let_have Mother never nothing; meaning: *Mother never treats herself to anything pleasant)*,

2. Coreference with relative elements. Cf.

   Za informační dálnici se považuje světová telekomunikační síť, po níž lze přenášet zvuk, data i obraz a která tak otevírá přístup k množství informatických služeb. (=...a net which makes it possible to transfer sound, data...)

   Here, the relative expression *níž (=which)* corefers with the noun *síť* (=*net*) modified by the dependent relative clause,

3. Coreference with verbal modifications that have a dual dependency (For more details see Mikulová et al. 2005),

4. Control. In the following example, the Actor of the infinitive *vyhlížet (=look like*) is controlled by the Actor of the verb *začít (=begin)*. The controller is the Actor of the control verb *začít.* The controllee is the Actor (subject) of the infinitive *vyhlížet.*

   Pokud dámy postupují podobně, {#PersPron.ACT} začnou brzy {#Cor.ACT} vyhlížet jako pánové. (=If ladies do the same they soon start to look like gentlemen)

5. Quasi-control (a specific grammatical coreference relation that can be found with multi-word predicates the dependent part of which is a noun with valency requirements). For more details see Mikulová et al. 2005.

6. Coreference in constructions with reciprocity. Cf.

Sultáni se vystřídali {#Rcp.PAT} na trůnu. (=lit. Sultans REFL changed on throne)

For a more detailed description of the types of grammatical coreference, see Mikulová et al. 2005.

4.2 Pronominal Textual coreference

Pronominal textual coreference is annotated in the following groups:

1. 3rd person personal and possessive pronouns; 1st and 2nd persons are excepted. (In the tectogrammatical tree, personal and possessive pronouns have the single t-lemma #PersPron.)

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.)

2. The demonstrative pronouns *ten, ta, to (=that)*.

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.)

3. With textual ellipsis, where a new node with the t-lemma substitute #PersPron is added to the tectogrammatical tree (if the added node represents a pronoun in the 1st or 2nd person textual coreference is not identified).

Myslíte, že rozhodnutí NATO, zda se {#PersPron} rozšíří, či nikoli, bude záviset na postoji Ruska? (=Do you think that NATO' s decision whether it {in Czech elided} will expand or not will depend on Russia' s attitude?)

5 Elements to be annotated

In this Section, we describe elements which are subject to annotation for coreferential and bridging relations. Our classification is based on the part-of-speech classification, using the terminology used for annotation of tectogrammatical level in Mikulová et al. 2005 (5.1), and the ability of elements to refer (5.2).

5.1 Part-of-speech classification

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 (see 5.1.1.4).

As for the difference between coreferential and bridging relations, it concerns only the number of relations, thus, we do not take it into account in this chapter.

The extended coreferential and bridging relations are to be marked between elements of the following categories:

- Complex nodes in the anaphoric position - nodes representing autosemantic lexical units, pronouns, ellipses, etc. (see 5.1.1),
- Quasi-complex nodes in the anaphoric position (nodes representing punctuation marks, non-alphabetical and non-numerical symbols),
- Paratactic structure root nodes in the anaphoric position (5.1.2),
- List structure root nodes in the anaphoric position (5.1.3).

5.1.1 Complex nodes in the anaphoric position

There are 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.

5.1.1.1 Semantic nouns in the anaphoric position

Semantic nouns are the most frequent subjects for coreference and bridging annotation. Semantic nouns are further divided into six groups. Nouns from all of these groups can be annotated for coreference and bridging relations. The following types of semantic nouns can be recognized:

1. Denominating semantic nouns (traditional nouns like *otec (=father), Marta* and possessive adjectives represented by the t-lemma of the corresponding nouns). In the tectogrammatical structure, they have the attribute sempos = n.denot.

Example:

> *a. Tímto faktorem je podnikatel – inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk, ani ztrátu. b. Podnikatelova* {coref_text, to „podnikatel“ in *a.*} *odměna, zisk, má však svůj původ nikoliv ve fungování, ale v rozbití stacionárního systému.* (= *a. This factor is enterpreneur-innovator, who... b. The enterpreneur's profit...*)

2. Denominating semantic nouns with which the negation is represented separately. These are denominating semantic verbal nouns ending in *-ní / -tí* (*hlasování* (=*voting*)) or *–ost* (*nezralost* (=*immaturity*)). In the tectogrammatical structure, they have the attribute sempos = n.denot. neg.

3. Definite pronominal semantic nouns: demonstratives. This group includes demonstrative pronouns in the positions of syntactic nouns. These are mainly demonstratives present at the surface structure (e.g.: *Ti už nepřijdou, O tohle mi nejde* (=*These will not come again; this is not the point*), also *tamten, onen, tenhleten* (=*that, this* etc.). Coreference of these kinds of pronouns had been already annotated as part of the pronoun textual coreference annotation (Kučová et al. 2003, Mikulová et al. 2005). In the tectogrammatical structure, they have the attribute sempos = n.pron.def.demon.

4. Definite pronominal semantic nouns: personal pronouns. This subgroup consists of all personal pronouns and their possessive counterparts (e.g.: *já, můj* (=*I, my*)), including the reflexives (*se / si, svůj*). All pronouns are represented by a single t-lemma: #PersPron, both, if they are present at the surface level or newly established in the tectogrammatical structure. In the tectogrammatical structure, pronominal semantic nouns have the attribute sempos = n.pron.def.pers.

Coreference of personal pronouns in the third person had been already annotated as part of grammatical coreference and pronoun textual coreference annotation. At the present stage of annotating extended textual coreference and bridging relations, personal pronouns (nodes with the t-lemma #PersPron) are included in the annotation mainly as antecedents of coreferential relations. The existing original coreferential chains are continued by newly established relations.

Fig. 1 shows the coreference chain (*Petr – on – Petr – on* (= *Peter – he – Peter – he*)), where the original pronominal coreference (*Peter – on* (= *Peter – he*)) is continued by extended nominal coreference (*on – Petr* (= *he – Peter*))

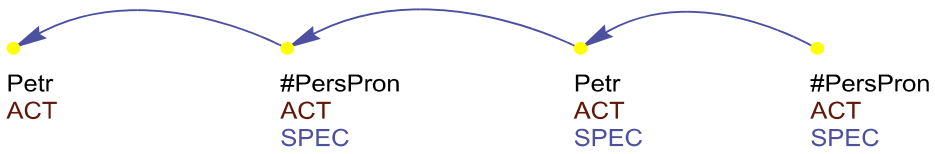

Fig. 1: Continuing coreferential chains

Coreference is not marked when a node represents a pronoun in the first or second person.

5. Indefinite pronominal semantic nouns. This group includes relative pronouns *kdo, co, který / jenž (=who, which) jaký (=which)* that are in the position of a syntactic noun (e.g. *Knihu, kterou si přál, nemohla sehnat* (= lit. *She couldn't find a book, which he wanted*, but not the cases of *Which book did he want?*)) and some of their derivatives, i.e. indefinite pronouns (e.g.: *někdo, některý* (=*somebody, some*)), negative pronouns (*nikdo* (=*none*)) and totalizers (*každý, všechen* (=*each, all*)). In the tectogrammatical structure, indefinite pronominal semantic nouns have the attribute sempos = n.pron.def.pers.

Relative pronouns have been annotated for coreference within the pronominal coreference annotation. In the annotation of extended textual coreference, they are annotated mainly as antecedents of coreferential and bridging relations. The rest of the listed pronouns had been annotated during the extended textual coreference stage. Cf.

> *a. X daroval Y počítače, kopírky apod. b. Vše* {bridging to „počítače", „kopírky" in *a.*} *v hodnotě 1 milión.* (= a. *X gave Y a computer, xerox machines etc. b. All of this was worth 1 million*.)

6. Numerals in the position of syntactic nouns (e.g. *Vybrali tři.* (=*lit. (They) chose three*), but not in *pět knih* (=*five books*))

5.1.1.2 Semantic adjectives in the anaphoric position

Adjectives are not subject to annotation of coreference and bridging relations, except for the following cases:

1. Adjectives refer to named entities. In such cases, adjectives are marked in the same way as nouns, from which they are derived. Cf.

Přijel do Prahy a pražská atmosféra se mu zdála celkem neformální. (= He arrived in Prague and found the Prague atmosphere quite casual.);

2. Adjectives with possessive meaning (*palácový* (= *lit. castle's*) – *palác* (*castle*), *dětský* (*children's*) in such cases as *dětská mysl /children's ideas*));

3. Adjectives with a demonstrative meaning (*tamní* (= *local*, lit. *from there*), *zdejší* (= *local*, lit.*from here*)).

We do NOT annotate:

1. Denominating semantic adjectives (*červený* (= *red*), *vlastní* (= *personal*), etc.),

2. Demonstratives in the position of semantic adjectives (ten dům se nám nelíbil (= *We didn't like that house.*)),

3. Indefinite pronouns in the position of semantic adjectives (*Kterou knihu si přál?* (=*Which book did he want to have?*)),

4. Numerals in the position of semantic adjectives (*Koupil pět knih* (= *He bought five books*)).

5.1.1.3 Semantic adverbs in the anaphoric position

Out of all groups of semantic adverbs (see Mikulová et al.. 2005), only definite pronominal semantic adverbs are annotated for coreference and bridging relations. These are definite demonstrative and identifying pronominal adverbs (e.g.: *tady, tam, teď, tak, tamtéž* (=*here, there, now, so, at_the_same_place*)) and their derivations (e.g.: *tudy* (=*this_way*) is derived from *tady* (=*here*), *odteď* (=*from_now*) is derived from *teď* (=*now*)).

5.1.1.4 Semantic verbs as members of a coreferential relation

Verbs are not annotated for coreference and bridging relation 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.

a. 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...*)

5.1.2 Paratactic structure root nodes in the anaphoric position

Root nodes of paratactic structures may be conjunctions used with coordination and apposition, e.g.: *a (=and), ale (=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 (e.g.: +, *krát (=times), od_ do (=from-to)*). Paratactic structure root nodes are common as coreferring and coreferred elements.

When choosing the antecedent by annotating coreference in sentences with coordination and apposiotion structures, annotation to the whole structure, i.e. technically to a paratactic structure root node, is preferred.

Examples:

a. Zápočty mezi podniky úspěšné. b. Úspěchem skončil podle včerejšího vyjádření ministra průmyslu Vladimíra Dlouhého dvouměsíční cyklus zápočtů závazků a pohledávek českých a {coref_text to „podnik" in a.} *slovenských podniků*. (= *a. Checks in companies were successful. b. According to yesterday's statement by the minister Vladimir Dlouhy, the two-month checks of Czech and* {coref_text from the conjunction to „podnik" in a.}*Slovak firms ended successfully.*) See also Fig. 2:

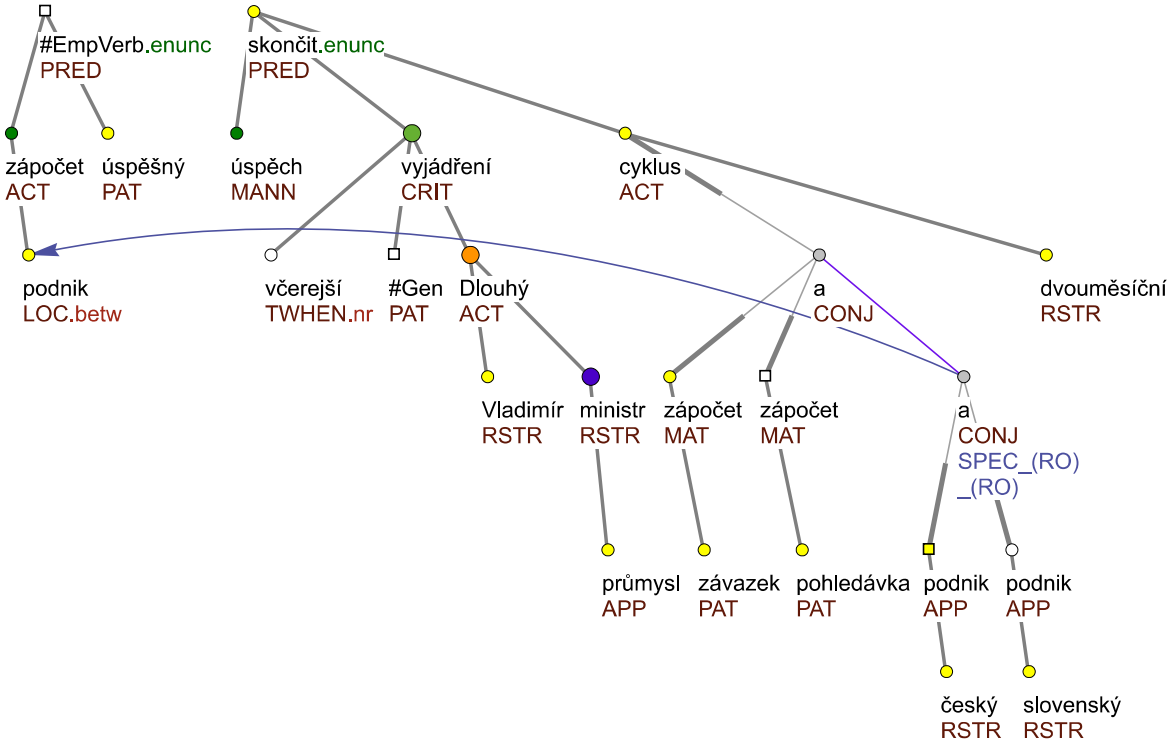

Fig. 2: Root nodes of paratactic structures in the anaphoric position

a. Zastihli jsme tady pouze jediného vychovatele – pana Fuchse. b. Podle pana Fuchse {coref_text, type = SPEC to #Dash, not to „pan Fuchse“} *nejde o žádné kriminálníky ani delikventní mládež.* (= *We met only one teacher, Mr. Fuchs. According to Mr. Fuchs....*). See also Fig. 3:

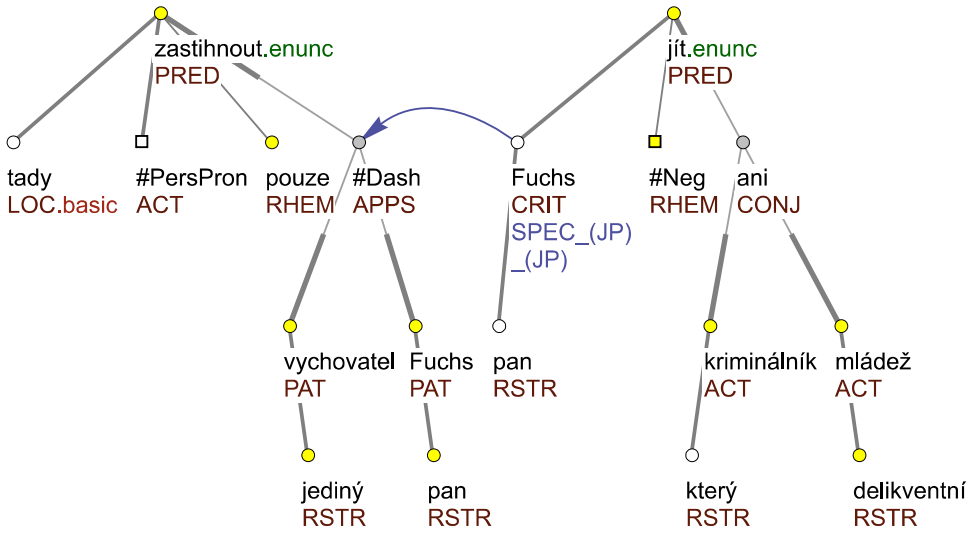

Fig. 3: Root nodes of paratactic structures in the antecedent position

5.1.3 List structure root nodes in the anaphoric position

Root nodes of list structures are nodes assigned the t-lemmas #Idph or #Forn. The function of these nodes is to assemble separate nodes into a list (structure). They have no counterpart in the surface structure of a sentence. They represent roots of structures, meaning mainly names e.g. of books or musical compositions (t-lemma `#Idph`), or if they put together words of a foreign phrase (t-lemma `#Forn`). When referring to these structures, coreference and bridging arrows always mark the list structure root nodes, i.e. nodes with t-lemmas `#Idph` or `#Forn.`

Examples:

a. První stěžejní prací dr. Svobody a jeho spolupracovníků byl reléový počítač SAPO. b. Pracovali na něm od roku 1950, do zkušebního provozu byl však `#Idph.PAT` {coref_text to "na něm"} *SAPO.*`ID` *uveden – vzhledem k obtížím při získávání součástek i k informační bariéře – až roku 1957. (= The first impediment to Dr. Svoboda's and his colleague's work was the SAPO computer. b. They had worked on it since 1950, while the SAPO {coref_text to "on it"}entered the testing phase only in 1957.).*

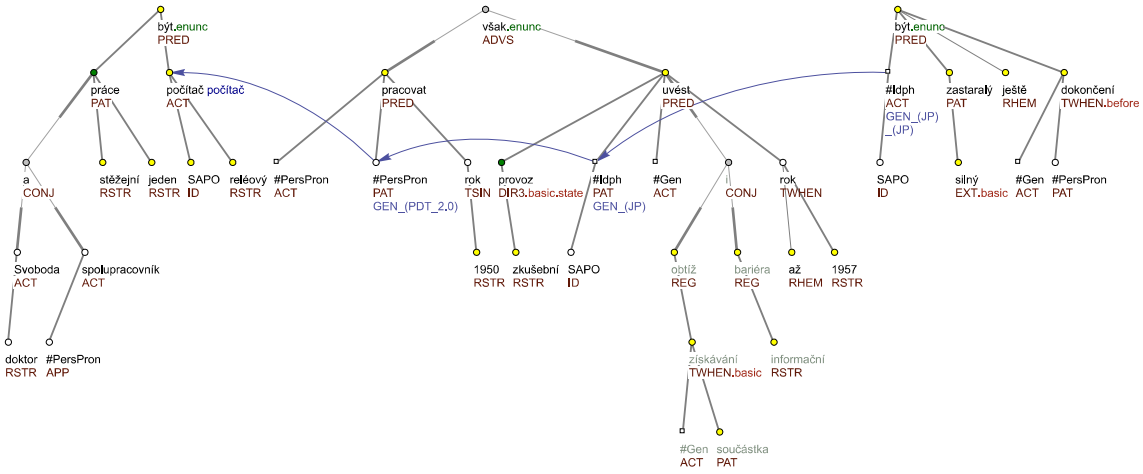

Fig. 4: List structure root nodes in the anaphoric position

a. `#Forn.ACT` *TTI.FPHR Therm.FPHR dodával stále vodoměry nedělené. b.* `Firmě` *dlouho trvalo, než* `#PersPron.ACT` *prosadila u německého producenta dělení vodoměrů. c. Jakmile* `#Forn.ACT` {coref_text to `#PersPron` in b} *TTI.FPHR Therm.FPHR začala dodávat dělené vodoměry, opět získala dřívější pozice na trhu. (= Therm used to deliver water measurer as a whole. b. It took the company a long time*

before it convinced the German producer to divide the water measurer. c. As soon as Therm began to deliver water measurements in parts...)

5.2 Referring and non-referring noun phrases

In the annotation of extended textual coreference, we distinguish between referring and non-referring NPs. Non-referring NPs are not to be annotated. The following noun phrases 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 je programátor.* (*= Petr is a programmer)* is NOT annotated as coreference. In the same way, coreference is not marked in identification structures (e.g. *Petr je právě ten zedník, který nám dělal koupelnu.(= 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 *norms* and the coordinative construction *regulations, bans and sanctions* in *Právo je souhrnem norem, to jest předpisů, zákazů a sankcí.* (*=lit. Law is a collection (of) norms, that is regulations, bans and sanctions*)), see the following example and Fig. 5.

Právo je souhrnem norem, to jest předpisů, zákazů a sankcí. (=lit. Law is a collection (of) norms, that is regulations, bans and sanctions)

3. Identifying expressions, which are represented as identification structures (in the tectogrammatical structure, they have the functor ID).

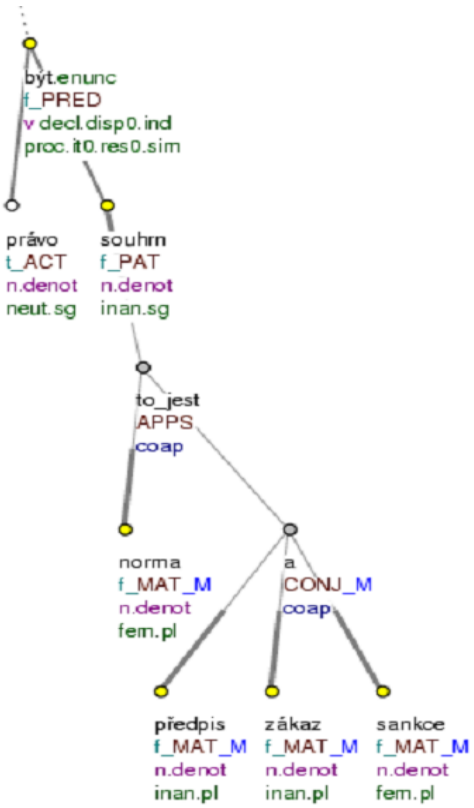

Fig. 5: Referring and non-referring noun phrases

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

   a. Americký index obchodní důvěry odbytu a zaměstnanosti v příštích šesti měsících se v srpnu snížil na 49,9 bodu, z 56,4 bodu {no coreference to „bod“} *v červnu. b. V dubnu byla jeho hodnota rovněž 49,9 bodu* {no coreference to „bod“}. *(= a. The American index of trade... and employment decreased in the last six months...., from 56.4 points in June to 49.9 points* {no coreference to „point“}*in August. b. In April, its value was also at 49.9 points* {no coreference to „point“}.*)*

Other NPs are considered to be referring and they may be annotated for coreference and bridging relations in PDT.

6 Annotation principles and preferences

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 (see 11.2.2.3). The chain principle does not concern bridging relations.

Principle of the maximum length of coreferential chains. This principle, similar to the chain principle, concerns only the cases of textual coreference. It says that in case of a multiple choice, we prefer to continue the existing coreference chain, rather than to begin a new one. To meet this principle, grammatical coreferential chains (already annotated in PDT) are being continued by textual ones, and similarly, the already annotated textual coreferential chains are continued by currently annotated non-pronominal links.

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. So, unlike MUC, we do not annotate predication and apposition relations.

Also bridging relations are not to be annotated if the anaphor is a direct child of its antecedent in the tectogrammatical tree, and it has some of the predefined labels for the valency relations (functors), such as PAT(iens), AUTH(or), APP(urtenance), etc.. So, for example, the relation between *strop* (*ceiling*) and *místnost* (*room*) in the phrase *strop této místnosti (= (the ceiling of the room)* is not annotated, because in the tectogrammatical tree, the node *místnost* (*room*) has the functor APP, being the direct child of the node *strop* (*ceiling*). Further details see in 9.2.

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, e.g.

> *"Duha?" Kněz přiložil prst k tomu slovu* {bridging to "duha", type ANAF}, *aby nezapomněl, kde skončil*. (="*Rainbow?" The priest put the finger on this word* {bridging to "duha", type ANAF}, *so that he didn't forget, where he stopped*.).

In a Slavonic language lacking the grammatical category of definiteness, we cannot afford to choose only definite NPs for anaphoric annotation (as it is done e.g. in the annotation schemas MATE (Poesio 2004), PoCoS (Chiarchos – Krasavina 2007) and AnCora-CO (Recasens – Martì 2010)), so we annotate all NPs that refer to the same entity. Non-coreferential anaphoric entities are annotated separately as a bridging relation (see 9).

Preference of coreference over bridging anaphora. The preference says that in case of multiple choice, we always prefer the textual coreference to a bridging relation. So having the following sequence of NPs: (a) *Mary* – (b) *John* – (c) *children of the class* – (d) *Mary and John*, we will annotate *Mary* in *d* as coreferential to *Mary* in *a*, rather than bridging_SUBSET to *chidren* in *c*, although this relation would be closer.

7 Extended textual coreference

Extended textual coreference is a case of coreferential relation, where anaphors are expressed neither by personal or demonstrative pronoun, nor are they elided. Here, we present the typology of coreferential relations, rules of annotation in the tectogrammatical structure in PDT and some problems we faced during the annotation process.

We do not annotate anaphoric relations in a restricted case, but we concentrate on marking the equivalence of referents of antecedent and anaphoric expressions.

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 do NOT annotate textual coreference in the following cases:

1. Relation between wh-words and replies to them (e.g. *Kdy mi zavoláš?* - *Večer.* (= *When will you call me?* - *In the evening.*))

2. Relation between personal pronouns in the first and second person. (e.g. *Budeš tam?* - *Ještě nevím.* (= *Will you be there? I don't know yet.*))

7.1 Typology of coreferential relations

In the tectogrammatical structure, referring expressions are not further classified into specifying and generic ones. Nevertheless, we assume generic NPs to have other anaphoric properties in the discourse. In addition, they result in greater ambiguity and are the cause of lower inter-annotator agreement. These were the reasons we decided to place them into a special

category of coreferential relations, thus forming separate coreferential chains of NPs with generic reference.

In the annotation of coreference, we distinguish between coreference of NPs with specifying reference and coreference of NPs with generic reference.

The information about the type of coreference is included in the value of the informal-type attribute inside the coref_text attribute. The informal-type attribute has two possible values and applies to every coreferring node in the tectogrammatical tree. The attribute values are listed in Table 1, "Values of the informal-type attribute".:

SPEC	textual coreference between NPs with specifying reference
GEN	textual coreference between NPs with generic reference

Table 1: Values of the informal-type attribute

In ambiguous cases with concrete nouns, the coreference type SPEC is preferred.

7.1.1 Coreference of NPs with specifying reference (value SPEC of the attribute informal-type)

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.

Examples of different textual coreference between specifying NPs:

– same t-lemmas

a. Jeho dojetí znásobila při vyhlašování přítomnost [...] pořadatelů soutěže – Českého manažerského centra v Čelákovicích. b. Na letošním ročníku soutěže {coref_text, type = SPEC to „soutěž" in a.} *se spolupodílí i Profit. (= His feelings were that applified during the announcement of the presence […] of the organizers of the competition.... b. Magzine profit is also taking part in this year's competition.)*

– different t-lemmas

a. Jak je dále v materiálu zdůrazněno, pozitivní posun v rozvoji malých a středních podniků byl umožněn především díky stabilnímu makroekonomickému prostředí, relativní legislativní stabilitě a státní politice podpory podnikatelských subjektů. b. Z dokumentu {coref_text, type = SPEC to „materiál“} *dále vyplývá, že v roce 1993 bylo celkově na podporu zejména malého a středního podnikání poskytnuto z rozpočtových prostředků více než 11 miliard korun. (= As emphasized in the materials, the positive trend in the development of small and medium size businesses was possible primarily thanks to... b. It is evident from the document that...)*

– different t-lemmas, a kind of hyperonymous relation

a. Usnesením vlády SR je koordinací všech akcí souvisejících se zajištěním certifikace dovážených potravinářských výrobků pověřen ÚNMS SR. b. Na tomto úřadě {coref_text, type = SPEC to „ÚNMS SR“} *lze získat i potřebné informace. (= According to a decree of the Slovak government, the coordination of all activities linked to the securing of certification for imported food items was entrusted to UNMS SR. b. Necessary information is provided by this institution.)*

– different subtrees with the same governing node

a. Nejvíce Ministerstvo financí .b. Nejvíce se na tom podílel resort Ministerstva financí ČR {coref_text, type = SPEC to „Ministerstvo financí“ in a.} *– a to formou daňových úlev ve výši zhruba 7,5 miliardy korun. (= a. Most of all Ministry of Finance did. b. The Ministry of Finance of the Czech Republic participated the most in this...)*

– non-specifying non-generic coreference

Například muž, který pracuje v nějakém velkém podniku, se zakouká do sekretářky ve stejném podniku {coref_text, type = SPEC to „podnik“} *a začnou se scházet v nějaké kavárničcce stranou od toho úřadu* {coref_text, type = SPEC „podnik“}. *(= For example, a man who works in a large company will fall in love for a secretary in the same company, and they will begin meeting up in some cafe away from that institution.)*

7.1.2 Coreference of NPs with generic reference (value GEN of the attribute informal-type):

Generally speaking, a noun phrase can be considered to be generic if it refers to a type or a prototype (e.g. *Univerzita není místo pro války.* (= *The university is no place to fight a war*.)).

Generic noun phrases are considered to be coreferential and they are correspondingly annotated if they refer to the same type of subjects to the same extent.

Textual coreference of type GEN is annotated in the following cases:

1. For generic noun phrases in singular and plural

– same t-lemmas

a. Nová striktní omezení vlády SR proti <u>českým exportérům</u>. b. Již několik dnů je všeobecně známo, že ochranářská opatření slovenské vlády proti <u>českým exportérům</u> {coref_text, type = GEN to „exportér“ in a.} *se dotýkají zejména oblasti obchodu s potravinami a zemědělskými produkty. (= The Slovak government's strict restrictions on <u>Czech exporters</u>. b. … protective measures of Slovakia's government against <u>Czech exporters</u>...)*

– pronominalisations and ellipsis

<u>Droga</u> je tedy tak účinná, že ten, kdo <u>ji</u> {coref_text, type = GEN to „droga“} *užívá, se snadno dostane do „pohody“ kouřením nebo šňupáním. (= The <u>drug</u> is so effective that the person who takes <u>it</u> can easy achieve the state of “coolness” by smoking or snorting.)*

– abbreviation – full expression

a. O odpočtu <u>DPH.</u> b. Podle novely zákona o <u>dani z přidané hodnoty</u> {coref_text, type = GEN to „DPH“} *se letos stanu plátcem daně. (= The subtraction of <u>VAT</u>. b. According to an amendment on <u>value added tax</u>, I will become a tax payer this year.)*

– between the nodes that depend on the words with the meaning of a “container”

a. V běžném vzorku sedmdesátých let byla pouze 3–4 procenta čisté <u>suroviny</u>. b. Nyní jsou k dostání balíčky obsahující až 80 procent čistého <u>heroinu</u> {coref_text, type = GEN to „surovina“}. *(= a. In an average sample from the seventies, there were only 3-4 percent of pure <u>raw material</u>. b. Currently, one can get packages containing up to 80 percent of pure <u>heroin</u>.),* see Fig. 6.

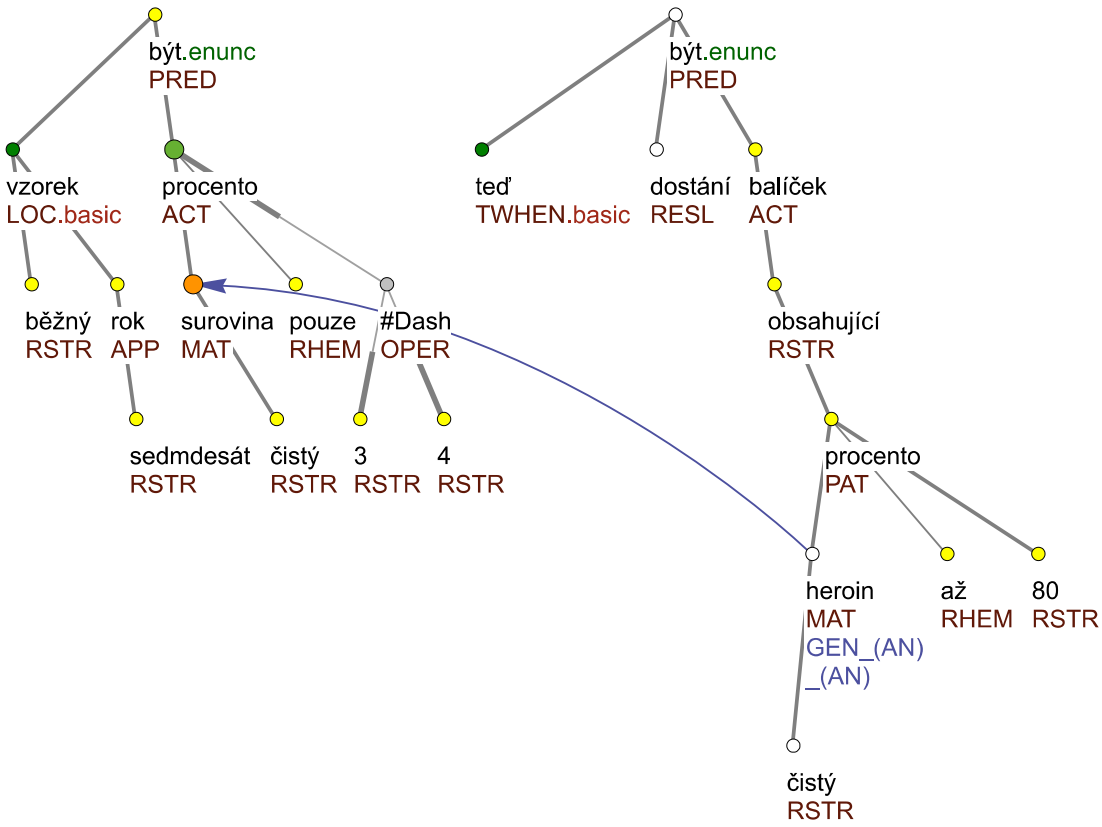

Fig. 6: Coreference of NPs with generic reference

2. For the majority of abstract nouns

Example:

Tímto faktorem je podnikatel-inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk {coref_text, type = GEN to „zisk“}, *ani ztrátu (= This factor is the enterpreneur-innovator, who is trying to gain profit, and hence, logically, cannot exist in a static state, where there is no profit or loss.)*

3. In case of non-generic non-specifying noun phrases, when antecedent and anaphoric noun phrases have the same t-lemmas and the same scope, but anaphoric NP doesn't have a determiner (else the SPEC coreference is annotated). Although this kind of relation does not supply text coherence very much, we still tend to mark this relation, also for the reason that the borderline between such cases and generics is not always clear.

Example:

a. Když si dítě bude přát, aby se o jeho problému nikdo z rodiny nebo školy nedozvěděl, musíme to respektovat, vysvětluje Jana Drtilová . […] b. Většinou se stává, že dítě ani

nechce, aby se rodina {coref_text, type = GEN to „rodina“ in a.} *dozvěděla, že se nám ozval. c. Linka by neměla rodinu {koreferenční vztah neanotujeme} nahrazovat, ale doplňovat. (= a. If a child desires that no one from the family or school would find out about his problems, we have to respect that, says Jana Drtilova. […] b. It is usually the case that the child does not even want for the family to know that he contacted us. c. The hotline should not replace the family, but to supplement it.)*

In case when the sentence a. is followed by the sentence c., coreference between a. and c. is not to be annotated. In a., NP *rodina* (= *family)* has non-generic non-specifying reference, and in c., this NP is generic. Even if taken in very wide sense, these NPs are not coreferential.

7.1.3 Borderline cases between SPEC and GEN coreference

In some cases, it is hard to define, if a noun phrase has a specifying or a generic reference. Mostly, both interpretations are possible. There are no firm rules for an unambiguous assignment of the types in these cases; the type is chosen on the basis of the available context and the annotator’s consideration.

Examples with the decisions annotators made in every specifying case:

a. Pracovníci zahraničních firem působících v České republice často tvrdí, že naši zaměstnanci nedosahují takových kvalit, jaké potřebují. [...] b. Jsou stesky na nekvalitní výkony našich lidí {coref_text, type = GEN to „naši zaměstnanci“ in a.} *oprávněné? (= Employees of foreign companies based in the Czech Republic often claim, that our workers do not have the necessary skills. […] (8) Is the criticism of the low productivity of our people fair?)*

a. U detergentu Toto jsme například řešili problém s udržením stálé kvality, protože jednotlivé partie byly nevyvážené. b. Investovali jsme dva miliony korun do nákupu pásových vah, zpřesnili dávkování a jakost pracího prášku {coref_text, type = SPEC to „detergent Toto“ in a.}*stabilizovali*. (*=For example, with the Toto detergent we face problems with maintaining consistent quality... b. We invested two million crowns... and stabilized the quality of the detergent.)*

a. Tím pádem máme problém se silniční daní. b. Váš problém vyřešila prosincová novela silniční daně {coref_text, type = GEN to „se silniční daní“ in a.}*, ve které bylo zrušeno osvobození od [silniční daně zaměstnaneckých vozidel]. (= a. As a result, we have a*

problem with the highway tax. b. Your problem was resolved by the December amendment of the highway tax, which ...)

7.1.4 Borderline between the GEN coreference and no relation

By generic NPs, we often don't know, if the case should be marked as coreference or it should not be annotated for coreference at all. Choosing between annotating or not annotating generic coreference, the following rule holds:

We do NOT annotate generic coreference in the following cases:

1. Noun phrases have different scope (i.e. they refer to different sets of objects, e.g. *žena (= women) – žena v 19. století (= women in 19th century)*)

2. If both a. and b. is true:

a. annotator is not sure about annotating generic coreference in this case,

b. the relation between NPs is not relevant for the coherence of the text.

Else, we annotate generic coreference.

Examples:

a. Kanaďané bodují b. V praxi jsem si potvrdil, že Kanaďané {coref_text, type = GEN to „Kanaďané“ in a.} *patří ke světové špičce také v leteckém průmyslu, poštovních službách a telekomunikacích, ve výrobě léků i dopravních zařízeních atd. (=a. Canadians are giving points b. Through personal experience I came to the conclusion that Canadians belong also to world's top aviation producers...)*

In this case, the coreferential relation has been annotated, both NPs are considered to be generics with the same scope.

a. Mnozí čeští podnikatelé si ke své škodě stále ještě neuvědomují, že peníze věnované na získání důležitých údajů se jim mnohonásobně vrátí. […] b. Naším cílem je, aby podnikatelé {no relation} *věděli o sobě navzájem*. (=*Many Czech entrepreneurs to their own disadvantage still do not realize that... b. Our goal is for entrepreneurs to know about each other.)*

Here, the coreferential relation was not annotated, the scope of the noun phrases *podnikatelé (=businessmen)* is not the same.

7.2 Textual coreference with special lexical groups

7.2.1 Coreference with abstract nouns

One of the most problematic areas in annotation of textual coreference in Czech is what to do with abstract nouns. On the scale of the referring ability, they are located somewhere between referring and non-referring nouns, but although the distinction between abstracts and concretes is basic, it is still very gradual and it's particularly hard to set formal criteria of what is and what is not subject to annotation of coreference.

In spite of all complications, we try to distinguish between specifying and generic coreference in annotation of textual coreference with abstract nouns. The annotating convention is the following:

If subjects to annotation have complements with specifying reference, or they have unambiguously specifying reference themselves, coreference between them is annotated as textual coreference, type SPEC. In case of even a little doubt annotate textual coreference, type GEN.

Examples:

We consider abstract noun to be specifying if they have specifying arguments, e.g. the abstract noun *ekonomika* (= *economics*) is specified by specifying *český (=Czech),* so the relation in the following example should be annotated as textual coreference, type SPEC.

> *a. Ve specifických podmínkách <u>české ekonomiky</u> růst nezaměstnanosti v letech 1991–1993 značně zaostal za poklesem HDP. […] b. Nejméně dvouprocentní růst <u>české ekonomiky</u>* {coref_text, type = SPEC to „ekonomika“ in a.} *již letos (1999). (=In the specific conditions of the <u>Czech economy</u> the growth of unemployment... b. This year (1999) at least a two percent growth of <u>the Czech economy</u>.)*

The relation between *kapitalismus (= capitalism)* in a. and b. and *socialismus (= socialism)* in b. and c. is annotated as textual coreference, type GEN, because the decision is ambiguous and the nouns don't have complements with specifying reference.

> *a. Dovozoval, že vývoj kapitalismu se historicky vyznačuje dvěma fázemi: Fází soutěžního kapitalismu a fází kapitalismu trustů. b. Schumpeter se ve svém posledním díle ptá: Který systém, kapitalismus* {coref_text, type = GEN to „kapitalismus“}, *či socialismus, bude určovat budoucnost lidstva? c. K údivu, úžasu či ohromení většiny svých kolegů odpovídá jednoznačně: Bude to socialismus* {coref_text, type = GEN to „socialismus“}. *(=a. He believed that the development of capitalism is historically apparent in two parts... b. Schumpeter asks in his last work: Which system, capitalism or socialism, will determine the future of man kind? c. To the surprise and amazement of his colleagues he answers: It will be socialism.)*

Cf. also the following:

> *a. Tímto faktorem je podnikatel-inovátor, který se snaží o zisk, a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk* {coref_text, type = GEN to „zisk“}, *ani ztrátu. b. Na konci tohoto difusního procesu se systém vrátí ke statické rovnováze, v níž nebudou opět ani zisky* {coref_text, type = GEN to „zisk“ in a.}, *ani ztráty. (=This factor is the enterpreneur-innovator, who is trying to gain profit, and hence, logically, cannot exist in a static state, where there is no profit or loss. b. At the end of this diffusionary process the system will return to a static equilibrium, in which neither profit or loss will exist once again.)*

7.2.2 Coreference with verbal nouns

The second problematic group for annotating textual coreference of both types is the group of verbal nouns. By verbal nouns, also specifying and generic reference is possible. Textual coreference with verbal nouns is annotated according to the following strategy:

1. If both members of the coreferential relation are verbal nouns with the concrete meaning, coreference between them is annotated in the same way as for other concrete nouns (see 7.1). Cf.:

> *a. Příslušnou rubriku najdete na 2. straně tiskopisu přiznání označenou jako položka 2 – Odpočet při změně režimu. b. Doklady k odpočtu se k přiznání* {coref_text, type = GEN to „přiznání“ in *a.*} *nepřikládají. (=a. You will find the relevant section on page 2 of the*

tax return form, marked as a Line 2... b. The documents are not to be attached to the tax return form.)

2. If the anaphoric element is a verbal noun with abstract meaning, there are four possibilities of annotation coreference in them:

a. If both verbal nouns are specifying, they refer to a specific situation and their possible arguments are coreferential, the relation between them is annotated as textual coreference, type SPEC. Cf.

a. Malé a střední podniky zvyšují svůj podíl na vyrobeném produktu i na zaměstnanosti a jejich počet neustále roste. b. Tuto skutečnost {coref_text, type = SPEC to „a“ in a.} *jednoznačně konstatuje ministr hospodářství Karel Dyba v analýze, kterou předložil vládě. (= a. Small and medium-size businesses increase their share of the product and of employment, and their numbers are growing. The Minister of Finance Karel Dyba unequivocally confirms this fact in...)*

b. If both verbal nouns are generic, or rather if their arguments are generic, the relation between them is annotated as textual coreference, type GEN. Cf.

a. Rychlé, avšak i bezpečné vypořádání. b. Rychlost vypořádání {coref_text, type = GEN to „vypořádání“ in a.} *burzovních obchodů v čase T + 3 odpovídá podle Jiřího Béra, ředitele Burzovního registru cenných papírů při Burze cenných papírů Praha potřebám. (=a. Fast, yet safe transaction b. The speed of transaction...)*

c. If both verbal nouns are specifying, but their arguments are not coreferential, coreferential relation between them is not annotated.

d. If one verbal noun is specifying and the second one is generic, coreferential relation between them is not annotated.

7.2.3 Coreference with named entities

Unlike other groups of nouns, in case of named entities, also adjectives derived from them may be annotated for coreference. However, de facto this exception holds mainly for adjectives derived from place names (*Praha (=Prague) - pražský (=lit. Prague)* etc.).

Textual coreference for place names has been automatically pre-annotated.

Examples:

a. Pouze z bývalé Šternberské konírny v přízemí křídla přiléhajícího k Thunovské uličce se stane konferenční (tiskový) sál. […] b. Když architekti zvažovali optimální propojení staré budovy sněmovny s novými domy, vsadili na tunel pod Thunovskou uličkou. {coref_text, type = SPEC to „Thunovské uličce“ in a.}*(=a. The conference hall will be constructed only in the old Sternberg stables on the ground floor of the wing adjacent to Thunoska street. b. […] they decided to build a tunnel under Thunovska street.)*

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 functor, coreferential relation is annotated to the governing node of the common noun. Cf. the following example and Fig. 7.

a. Právě na distribuci měřicích přístrojů se zaměřila ostravská firma TTI Therm. b. Ředitel Rudolf Kluziewicz (na snímku) založil firmu {coref_text, type = SPEC to „firma“ in a.} *spolu s partnery z Německa. (=a. It was the distribution of measuring instruments, the TTI Therm company from Ostrava decided to focus on. b. The director Rudolf Kluziewicz (in the photo) founded the company together with partners from Germany.)*

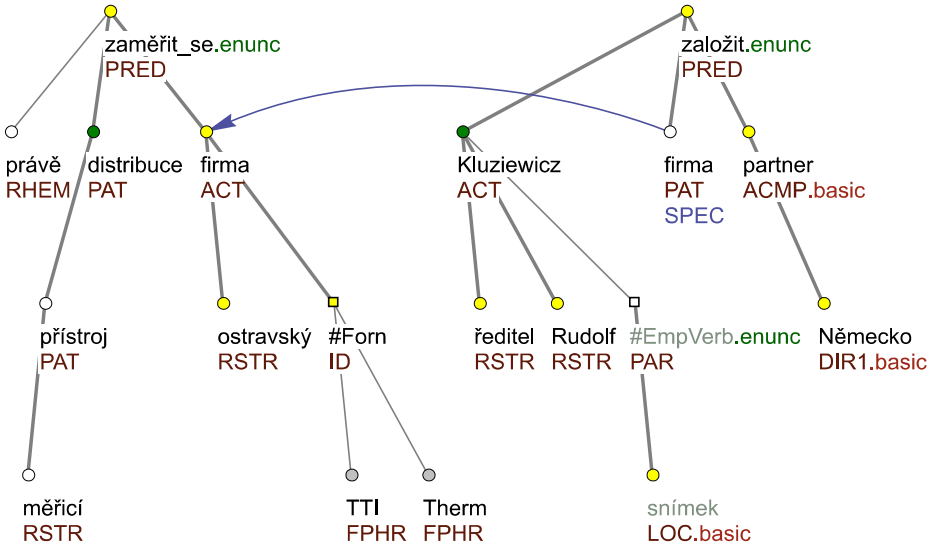

Fig. 7: Coreference with named entities

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 coreference relation between NPs *Vaclav Havel* illustrated on Fig. 8:

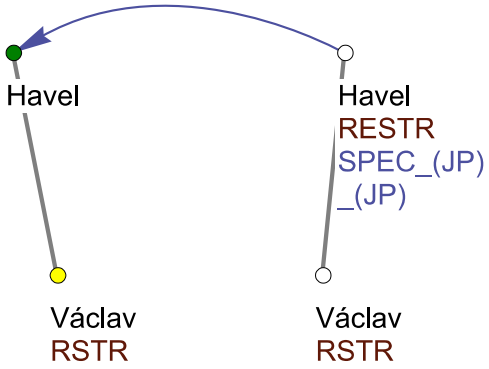

Fig. 8: Coreference with named entities

If automatically pre-annotated otherwise, it is manually corrected. Cf. Fig.9:

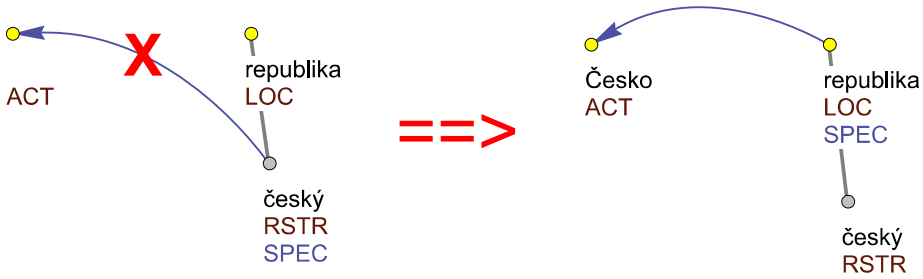

Fig. 9: Coreference with named entities

Proper names are not considered to be atomic, and nested mentions inside them may be annotated separately, but only in case if they are named entity themselves. So we annotate coreference for *ČR* (=Czech Republic) inside the phrase *Ústavní soud ČR* (= *the Law Court of the Czech Republic*) but we do NOT annotate coreference for *výzkum rodiny* (= *research on family life*) inside *Oddělení pro výzkum rodiny* (= *Department for research on family life*.)

8 Special types of textual coreference (coref_special)

Two special cases of (co)reference are annotated in PDT inside the group of textual coreference:

- references to situations or reality external to the text (coref_special, type exoph), see 8.1;
- references to a discourse segment consisting of more than one sentence (coref_special, type segm), see 8.2.

Both cases have been already annotated in the phase of pronoun coreference annotation (Kučová et al. 2003). Here, the extension of these types to the full noun phrases is presented.

8.1 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 noun phrases with demonstratives may be annotated for exophora.

Exophoric reference is annotated in the following cases:

1. Time and local deixis, e.g.

   Dokončeny by měly být do 31. prosince 1995, a to i přes jisté zdržení způsobené opožděným stěhováním nájemníků z domů čp. 8 a 518 do náhradních bytů na sídlišti Barrandov v <u>těchto dnech</u> {coref_special, type = exoph}*.(= It should be finished before 31 Dec 1995, despite some delays caused by the late departure of renters from houses no. 8 and 518 to replacement flats in Barrandov district <u>in these days</u> [meaning, in the recent days])*

   Plukovník StB Čadek udělal chybu, když poslechl doporučení novinářů a uchýlil se do předem připravených pozic. Neměl podlehnout panice a v klidu vyčkat věcí příštích. Vždyť v <u>téhle zemi</u> {coref_special, type = exoph} *se zatím žádnému skutečně velkému podvodníkovi nezkřivil ani vlas na hlavě. (= State Police Colonel Čadek made a mistake when he listened to the recommendations of journalists and retired to the previously prepared positions. He should not have succumbed to panic, he should rather wait for better things to come. Indeed, <u>in this country</u>, he didn't harm any really big cheater.)*

2. Deixis with pronominal adverbs, e.g.

   A <u>tu</u> {coref_special, type = exoph} *se dostáváme zpět k počátku tohoto textu . (= <u>With this</u>, we come back to the beginning of this text)*

3. Exophoric reference to the whole text, e.g.

Informace v tomto přehledu {coref_special, type = exoph} *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 showing 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), letos (=* lit. *this year), 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.

   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.

   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. ")

8.2 Reference to a segment

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 8.2.1),

- a noun phrase refers to a tree segment which cannot be technically separated (see 8.2.2)

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.

8.2.1 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 7.

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.

Cf. reference of *v tomto směru (=in that direction)* in the following example:

> *a. Celní unie bude sice existovat na papíře ještě dalších dvanáct měsíců, ale v praxi dostanou vzájemné vztahy punc tvrdosti mezinárodního obchodu. b. Poroste administrativa. c. Jistotu v tomto směru* {coref_special, type = segm} *dávají nejnovější kroky vlády SR, která se rozhodla zavést již před časem avizovanou desetiprocentní dovozní přirážku na zboží zahraniční provenience. (= a. A tax-free union will exist on paper for another 12 months, but in reality, the relationships will get a taste of the touch atmosphere of international trade. b. Administrative procedures will increase. c. The newest steps of the Slovak government are a source of confidence in this respect...)*

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

8.2.2 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 annotated as coref_special, type = segm.

> *a. Od 1. dubna nebude ÚNMS SR rozhodnutí české zkušebny potvrzovat. b. Tato funkce* {coref_special, type = segm} *přejde na příslušnou slovenskou zkušebnu, která bude vydávat na základě dodaných podkladů příslušné certifikáty*. (*=a. 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...*)

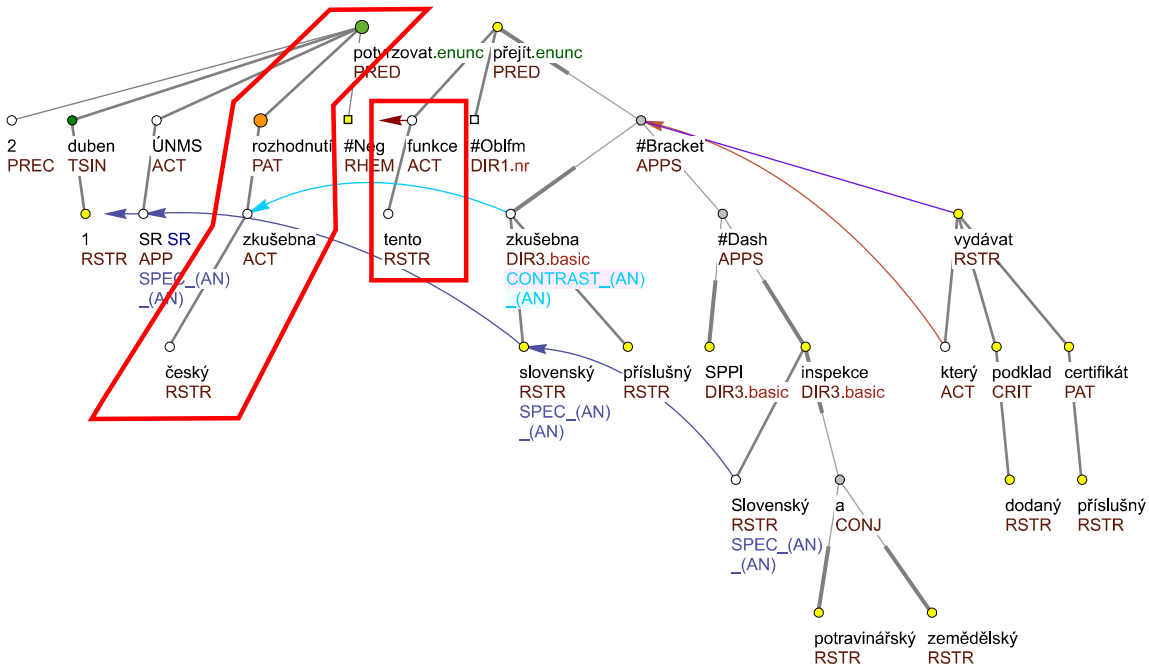

Fig. 10: Reference to a tree segment which cannot be technically separated

9 Bridging Relations

Two expressions are linked by a bridging relation (in other related projects also called bridging anaphora, indirect anaphora) if they stay in a specific non-coreferential semantic, lexical or conceptual relation. The restrictive sub-types of bridging relations may be chosen differently according to the aims of the given project.

As we assume that all elements of a coreferential chain in PDT are equivalent, a bridging relation connects the anaphoric NP with the closest node of the coreferring chain of the antecedent (including grammatical coreference). Cf. the following example and Fig. 11, where the anaphoric NP *meniskus (=meniscus)* is connected to the relative pronoun *který (= who)*, which is the nearest node in the coreferential chain *Šmicer - který*:

> *Šestigólovou výhru Slavie nad Brnem ve 24. ligovém kole sledoval z tribuny i útočník domácích Vladimír Šmicer, který* {coref_gram to "Šmicer"} *se ve čtvrtek podrobil půlhodinové operaci menisku* {bridging WHOLE_PART to „který"}. (=*Also Vladimir Smicer, who was operated on meniscus on Thursday, was watching the 6-goal Slavia's win over Brno by the 24 league round ...*)

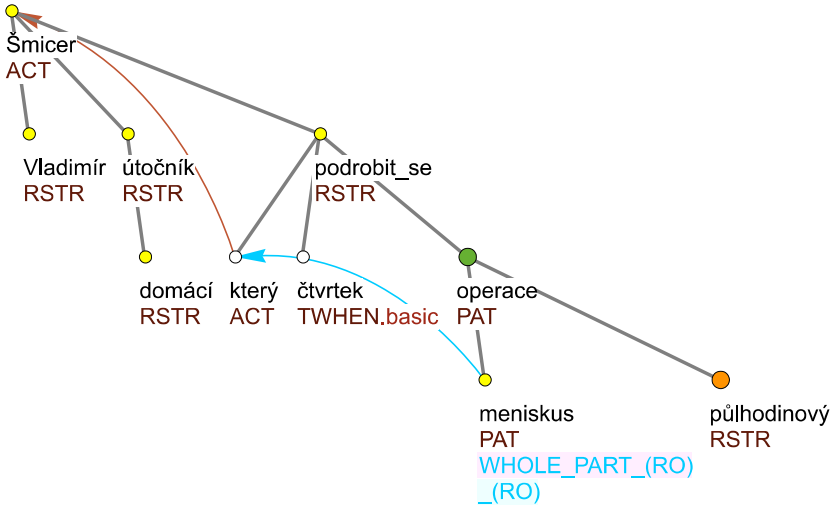

Fig. 11: Bridging relations

Bridging relations in PDT are annotated only between nominal expressions, no verbs are considered as anchors.

In our project, we annotate the following bridging relations:

- meronymical relation between a part and a whole (with subtypes PART_WHOLE and WHOLE_PART), see 9.1.1;
- the relation between set and its subsets/elements of the set (with subtypes SUB_SET and SET_SUB), see 9.1.2;
- the relation between an entity and a singular function on this entity (with subtypes P_FUNCT and FUNCT_P), see 9.1.3;
- the relation between coherension-relevant discourse opposites (type CONTRAST), see 9.1.4;
- non-cospecifying explicit anaphoric relation (type ANAF), see 9.1.5;
- further underspecified group REST, see 9.1.6.

9.1.1 Meronymical relation between a part and a whole (PART: PART_WHOLE and WHOLE_PART)

The meronymical relation between a part and a whole is one of the basic bridging relations and it is commonly agreed upon (Clark 1977, Muller – Strube 2001, Chiarchos – Krasavina 2007).

This relation has two directions – the type "PART_WHOLE" is used for the case when the antecedent of the anaphoric NP corresponds to the whole of which the anaphor is a part and "WHOLE_PART" for the opposite.

Prototypical examples are:

pokoj (=room) – strop (= ceiling),

ruka (=hand) – prst (=finger),

město (=town) – ulice (=street),

týden (=week) - pondelí (= Monday), etc.

In PDT, the PART bridging relation is annotated in the following cases:

1. By expressions referring to places: states, regions, towns, streets etc. (type *Německo* (=*Germany*) – *Bavorsko* (=*Bavaria*)– *Mnichov* (=*Munich*), *město* (=*town*)– *ulice* (=*street*)). Cf.

> *a. Jejich vysílače dosud pokrývají signálem programu ČT 2 méně než polovinu území <u>republiky</u> {meaning „Czech Republic"}. b. Do rozdělení <u>federace</u> {bridging, type = PART_WHOLE, to „republika" in a.} totiž signál zajišťovaly vysílače v <u>SR</u>* {bridging, type = WHOLE_PART, to „federace"}. *(=a. ... less than a half of <u>the country</u> {meaning „Czech Republic"}. b. Before the division of <u>the federation</u>, the signal was provided by signal transmitters in the <u>Slovak Republic</u>.)*

2. In prototypical cases of inseparable parts, which cannot be understood as subsets (type *pokoj (=room) – strop (=ceiling), ruka (=hand) – prst (=finger)*). Cf.

> *a. Jednotlivá <u>studia</u> v apartmánech jsou vybavena <u>kuchyní</u>* {bridging, type = WHOLE_PART, to „studia"}, *takže je možná individuální příprava stravy. (=<u>Studio apartments</u> are equipped with <u>kitchens</u>, so everyone may prepare his food himself.)*

3. By references to time distances. Cf.

a. Dělal jsem bez přestávky celé týdny, často v noci {bridging, type = WHOLE_PART, to „týden“}. *(=I worked nonstop for weeks, often at night.)*

9.1.1.1 Borderline cases with the PART bridging relation

Border with no relation. If a potential expression is not a part of a place, but it is located there, the bridging relation PART is not annotated. Cf.

a. V Mnichově jsou muzea {no bridging relation} *a galerie* {no bridging relation} *se vzácnými obrazy* {no bridging relation}, *částečně jsem je navštívil a zhlédl překrásný královský zámek Nymphenburg* {no bridging relation}. *(=a. In Munich, there are museums and galleries with rare paintings, I've visited some of them and I've also sen the beautiful Royal Castle Nymphenburg.)*

Border with the FUNCT bridging relation. The relation between place and a part of this place, and this part has functional interpretation. These cases are ambiguous, the choice of the relation depends on the decision of the annotator.

a. Slovensko po několika měsících diskusí devalvovalo svou měnu o deset procent. b. Je to jistě rozumné opatření. c. O tom, že současná hodnota Sk je neudržitelná, pochyboval totiž jen málokdo. d. Spíše je otázkou, zda Bratislava {bridging to “Slovensko”} *nepřistoupila k akci poněkud pozdě. (=After several months of discussions, Slovakia devalued its currency by ten percent. It is certainly a reasonable measure. Nobody doubted that the current value of Slovakia is untenable. The question is whether Bratislava was not somewhat late with this decision.)*

Border with the SUBSET bridging relation. There is a large borderline area with the SUBSET relation. For details see 9.1.2.2.

9.1.2 The relation between set and its subsets/elements of the set (SUBSET: SUB_SET and SET_SUB)

This relation has two directions – the type “SUB_SET” is used for the case when the antecedent of the anaphoric NP corresponds to a subset or an element of the set of which the anaphor is a set, and “SET_SUB” is used for the opposite.

Prototypical examples are:

nápoje (=drinks) – pivo (=beer) – limonáda (=lemonade) – minerálka (=soda) – cola (= coca);

motýli (=butterflies) – červení (= red ones) – bílí (=white ones);

semináře (= seminars) – první seminář (= first seminar) – poslední seminář (= last seminar)

Graphically, the SUBSET relation may be visualized as follows:

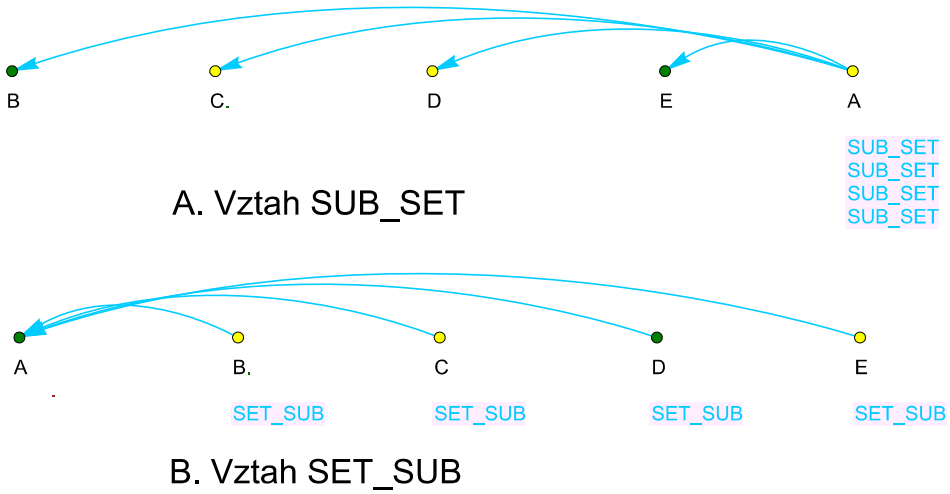

Fig. 12: Reference to a tree segment which cannot be technically separated

Examples:

a. Nejvíce se na tom podílel resort Ministerstva financí ČR – a to formou daňových úlev ve výši zhruba 7,5 miliardy korun. b. Další podpora byla poskytnuta Ministerstvem zemědělství – přibližně 2,8 miliardy korun, Ministerstvem práce a sociálních věcí – kolem 178 milionů korun a Ministerstvem hospodářství – asi 1,2 miliardy korun. c. V rámci rozpočtové podpory poskytují ministerstva {bridging, type = SUB_SET to „Ministerstvo financí ČR“ in a., „Ministerstvo zemědělství“, „Ministerstvo práce a sociálních věcí“ and „Ministerstvo hospodářství“ in b.} *malým a středním podnikatelům zvýhodněné informační služby a poradenskou činnost buď přímo, nebo prostřednictvím specializovaných institucí. (=The Ministry of Finance of Czech republic participated at this most of all … b. Further support was provided by the Ministry of Agriculture , it was approximately 2.8 billion, the Ministry of Labour and Social Affairs gave around 178 million crowns and the Ministry of economy gave about $ 1.2 billion. c. As the part of financial support, the ministries provide … information services … for small and medium-sized enterprises …)*

a. Pokud tedy zrovna nesedí na svém minikřesle v jednací síni, jsou poslanci nuceni pobývat buď ve svých klubech, nebo postávat či posedávat po chodbách. b. Nelze se pak ani divit, že část zákonodárců {bridging, type = SET_SUB, to „poslanec“} *zvolí příjemnější variantu a odchází úřadovat do suterénní restaurace zvané dolní sněmovna. (=a. If not sitting directly in their little chairs in the courtroom, deputies have to either stay in their clubs, or stand or sit around in the corridors. b. So one cannot be surprised that some lawmakers choose working in the basement restaurant ...)*

9.1.2.1 The SUBSET relation with abstract and verbal nouns

Annotating the SUBSET relation with abstract and verbal nouns appears to be quite a serious problem. This relation has a different meaning compared to the SUBSET relation of specifying nouns. However, the SUBSET-like relations between abstract and verbal nouns may be relevant for cohesion. For this reason, such kind of relation may be annotated as bridging SUBSET in PDT. The most frequent types are the following:

1. The relation “generic expression – a specifying example”. Cf.

   a. Nový VW Golf je vybaven motorem o síle... b. Dostali jsme možnost se novým golfem {bridging, type = SET_SUB to „Golf“ in a.} *projet.(=a. The new VW Golf is equipped with an engine power ... b. We had an opportunity to ride a new golf.)*

2. The relation “category – sub-category”. Cf.

   a. I když konzervativní Anglie jeho čin odsoudila, guma se zde chytila a Británie se pro žvýkačku stala bránou do Evropy. b. Ještě jeden milník si zaslouží zmínku – zrod bublinové žvýkačky {bridging, type = SET_SUB, to „žvýkačka“ in a.} *(=a. Although conservative England did not accept it, ... for the gum, Britain has become the gateway to Europe. b. Another milestone is worth mentioning, that is the birth of a bubble gum.)*

3. The relation between a set of specifying objects and a non-specific element. Cf.

   a. [volontéři] Absolvovali školení v první pomoci pro člověka v nouzi . […] b. Když dítě zavolá, dostane buď radu hned, nebo si s ním volontér {bridging, type = SET_SUB to „volontér“ in a.} *domluví další hovor. (=a. The [volunteers] have been trained in first aid for people in need. [...] B. When a child calls, it will get get an advice immediately, or a volunteer will arrange a meeting with him.)*

9.1.2.2 Borderline cases with the SUBSET bridging relation

Border with the PART bridging relation. In some cases, the distinction between "part-of" and SUBSET groups is quite problematic, so that the only reason to decide for the type of a bridging relation is the countability of corresponding nouns. Ambiguity is more frequent by generics, abstract and verbal nouns, but it can also appear by specifying expressions.

For the time being, the instruction for such type of ambiguities is to annotate the type PART only in clear cases of non-separable parts. If some doubts on ambiguity exist, annotate the SUBSET relation.

Examples:

> *Revidoval text Prezidentské adresy. Poslední věta* [bridging to "text", type WHOLE_PART or SET_SUB], *kterou v životě napsal, zněla ...* (= *He edited the text of President's address. The last sentence was...*)
>
> *Ročně by tedy zaplatila na pojistném, včetně úrazového připojištění* {bridging SUBSET nebo PART to „pojistné"}, *4 104 korun*. (=*Thus, she would pay the insurance every year, including the accident insurance, that is 4,104 crowns.*)

Border with textual coreference, type GEN. The cases of ambiguity between textual coreference of type GEN and the bridging SUBSET relation are quite often in texts with many generic noun phrases, and they are mainly caused by different depth of the referential interpretation. Cf.

> *V Plzni je stánkařům k dispozici tržnice pronajatá soukromé firmě Bratři Flaxové. Prodává se také na náměstí, od prodejců vybírá poplatek každé ráno správce tržiště*. (=*In Pilsen, the market … is available to kiosk owners . It is also marketed in the square, the custom duties are collected from the sellers every morning...*)

In accordance with our principles of the preference of coreference relations (see 6), in such cases textual coreference is preferred.

Border with coref_special, type = segm. In the following example, the interpretation of coreference of pronoun *to* is ambiguous – it may refer either to both preceding sentences or to the set of noun phrases {*patřičný rytmus* (=*proper rhythm*), *režim* (=*routine*), *vysoké využití podhorských pastvin* (=*high utilization of foothill grasslands*), *nejkvalitnější stáda* (=*finest herd*), …}:

a. Spolupráce by měla dostat patřičný rytmus, režim. b. Vysoké využití podhorských pastvin, nejkvalitnější stáda.c. To jsou předpoklady pro výrobu kvalitních potravin. (*=a. The cooperation should get the proper rhythm, its routine. b. The use of high foothill grasslands and finest herd.c. All these are the prerequisites for the production of high quality food.*)

In such cases, the SUBSET bridging relation is always preferred.

Border with the FUNCT bridging relation. For details see 9.1.3.1.

9.1.3 The relation between an entity and a singular function on this entity (FUNCT: P_FUNCT and FUNCT_P)

The FUNCT bridging relation is annotated between two entities when one entity has a singular function on another entity.

This relation has two directions – the type "FUNCT_P" is used for the case when the antecedent corresponds to a function on the anaphor which is in the anaphoric position, and "P_FUNCT" for the opposite.

Prototypical examples are:

trenér (= trainer) – mužstvo (*= team),*

premiér (*= prime minister*) *– vláda* (*= government*),

firma (*= company*) *– ředitel* (*= director),*

akce (*= event*) *– organizátor* (*= organizer*).

Examples:

a. Na přímou podporu podnikání vydá letos stát přibližně 1,8 procenta hrubého domácího produktu. b. Tuto skutečnost jednoznačně konstatuje ministr hospodářství Karel Dyba v analýze, kterou předložil vládě {bridging, type = P_FUNCT, to „stát"}. *(=a. The state will give about 1.8 percent of gross domestic product to direct business support this year. b. This fact is clearly statet by Economy Minister Karel Dyba in his analysis which he presented to the government .)*

a. Společnost zaměstnává přes dva tisíce zaměstnanců, to je po propuštění důchodců a brigádníků prakticky stejně jako před pěti lety. b. S generálním ředitelem Miloslavem Handlem {bridging, type = P_FUNCT, to „společnost" in *a.*} *jsme hovořili o tom, co se*

za těmito výsledky skrývá. (*=a. The company employs over two thousand employees, … . b. With the CEO Miloslav Handl, we talked about ...*)

9.1.3.1 Borderline cases with the FUNCT bridging relation

Border with the SET bridging relation. The distinction between types SUBSET and FUNCT is defined on the base of singularity of a singular function. For this reason, e.g. relation between *ministr (= minister)* and *vláda (= government)* is marked as SUBSET, while the relation *premiér (= prime minister) – vláda (= government)* is annotated as FUNCT.

S oznámeným snížením horní sazby daně z přidané hodnoty a daně z příjmu právnických osob o jedno procento vláda v zásadě souhlasí, a jak po jednání uvedl premiér {bridging, type = P_FUNCT, to „vláda“} *Václav Klaus, nelze v tomto bodě očekávat významnější změny. (=The government agrees … to the announced reduction of the top tax rate … and as the Prime Minister said ...)*

Border with no bridging relation. In some cases it is hard to decide, if the relation is still coherence-important and should be annotated as bridging or it should be omitted. If the case is ambiguous, it is up to the annotator to decide which interpretation is involved. The recommendation is rather not to mark clearly ambiguous cases at all.

Examples:

Rozhodování podat si žádost o osvojení dítěte není pro manžele vůbec snadnou záležitostí a může znamenat i konec manželství {no bridging relation}, kdy se jeden z partnerů s tímto zásadním obratem nedokáže vyrovnat. (=Deciding to make a request for adoption of a child is not an easy decision for spouses and it may mean the end of the marriage relationship, if one of the partners cannot accept the situation.)

Navíc mnoho nadaných studentů si vybralo po ukončení studií právě mecenáše své školy {no bridging relation} jako zaměstnavatele. (=Deciding to make a request for adoption of a child is not an easy decision for spouses and it may mean the end of the marriage relationship, if one of the partners cannot accept the situation.)

9.1.4 The relation between coherension-relevant discourse opposites (type CONTRAST)

The CONTRAST bridging relation is annotated between noun phrases standing in the relation of discourse opposites.

This relation has only one direction and it concerns only the noun phrases for which it is annotated, not the whole coreference chains.

There are no prototypical examples, the CONTRAST relations are figured out on the base of the context.

The CONTRAST relation is not really a bridging relation in the restricted sense, it could be rather labelled rhetorical or something like that. However, this kind of semantic dependence has a similar influence on the text cohesion as bridging relations. In addition, it supplements the similar kind of information in the topic-focus articulation annotation, where contrastive topic is marked, and the currently annotated contrast on the discourse level (Mladová – Zikánová – Hajičová 2008).

Examples:

a. Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 mil. USD, tj. téměř 2 % HDP. b. I když letos a {bridging, type = CONTRAST from the comma „a“ to „vloni“ in a.} *příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu, prognózujeme, že saldo přesto zůstane kladné ve výši 300–600 mil. USD ročně (1–1.6 % HDP). (=a. The balance of payments was estimated for about U.S. $ 600 million ... last year .... b. Although this and the next year there should be a slowdown in export growth …)*

a. Dnes, po rozdělení ČSFR, je jasné, že osud ČR bude stále více spojený s Německem a přes něj s Evropskou unií a osud Slovenska {bridging, type = CONTRAST to „osud ČR“ in a.} *s Ruskem.* (*=a. Nowadays, after the split of Czechoslovakia, it is clear that the fortune of the Czech Republic will become more associated with Germany, further with the European Union, while the fortune of Slovakia will be more associated with Russia.*)

9.1.4.1 A borderline case with the CONTRAST bridging relation

The most problematic bridging contrasts are relations between prepositional phrases, where nested noun phrases may be annotated for textual coreference (e.g. *před válkou (=before the war) – po válce (= after the war), v Praze (= in Prague) – kolem Prahy (= near Prague)*).

Border with the ANAF bridging relation. See details in 9.1.5.1.

9.1.5 Non-cospecifying explicit anaphoric relation (type ANAF)

In the non-cospecifying explicit anaphoric relation where the anaphor is marked with a demonstrative, bridging type ANAF is marked.

The bridging ANAF relation has one direction, it always refers back to the antecedent.

Prototypical examples look weak without context, e.g.

leden (= January) – ve stejném období loňského roku (= at the same time of the last year),

duha (= rainbow) – toto slovo (= this word),

Rakousko přepadlo Maďarsko (= Austria attacked Hungary) – v tu dobu (= at that time), etc.

Bridging relation ANAF is marked in the following cases:

1. Metalinguistic references, references to an antecedent expression, not to an extra-linguistic object.

Examples:

> *"Duha?" Kněz přiložil prst k tomu slovu* {bridging to "duha", type ANAF, *aby nezapomněl, kde skončil*. (="*Rainbow?" The priest put the finger on this word, so that he didn't forget, where he stopped.*)
>
> *a. Pavel Vondráček: Termín.PAT převýchova.ID znám pouze z nacistického a komunistického slovníku. b. Na převýchovu* {bridging ANAF to „termín"} *se, pokud vím, posílali ti, kteří měli podle těchto zrůdných režimů nevhodný původ. (=a. Pavel Vondracek: I know the term re-education only from the Nazi and Communist vocabulary. b. As far as I know, those people were sent to re-education, who were considered to have inappropriate origin by these monstrous regimes.)*

2. Anaphoric reference to the time, when the antecedent situation takes place, and the antecedent situation is expressed by a single sentence (else see 8.2.1).

Examples:

a. Tak jako každý Mexičan, i Santa Anna znal a občas žvýkal mízu sapodilly zvanou chicle, a tak se zrodil nápad pokusit se z chicle udělat náhražku kaučuku. b. Právě v té době {bridging, type = ANAF to the whole a.} *přihrála náhoda Santa Annovi do cesty Thomase Adamse, fotografa a především vynálezce všeho druhu. (=a. As every Mexican, Santa Anna also knew and occasionally a chewed chicle sap called Sapodilla, and thus the idea of trying to substitute for chicle to make rubber was born . b. At that time, Santa Anna had a chance to go …)*

a. Rozbití Varšavské smlouvy bylo jako odseknutí údů od těla. b. Od té doby {bridging, type = ANAF to „rozbití" in a.} *se toho mnoho neudělalo. (=The disintegration of the Compact of Warsaw was …. From that time, there was not much that was done.)*

3. Anaphoric reference to an object, which is similar to its antecedent in some characteristics. Usually, complements like *takový (= such), podobný (= similar), stejný (= the same)* are used with the noun phrase in the anaphoric position.

a. Nic nenasvědčuje tomu, že by parlamentní budova měla sloužit jiným než parlamentním účelům. b. Přesto se takové názory {bridging, type = ANAF to „sloužit" in a.} ozývají. (=a. There is no indication that the parliamentary building could serve for other purposes than parliamentary. b. However, one can hear such opinions.)

9.1.5.1 A borderline case with the ANAF bridging relation

Border with the CONTRAST bridging relation. This borderline takes place first of all in references to objects, which are similar to their antecedents in some characteristics. If the anaphoric noun phrase is specified by alternators (e.g. *jiný (= other)*) , the CONTRAST type is annotated. Else, the ANAF should be marked.

Examples:

… náklady nutného a neodkladného léčení v zahraničí... do výše nákladů spojných s takovým léčením na území ČR {bridging, type = ANAF to „léčení"}... *(=... the costs of necessary and urgent treatment abroad ... up to the amount of costs connected with such treatment in the Czech Republic)*

V Maďarsku je počet úmrtí v důsledku srdečních a cévních chorob větší než v kterékoliv jiné průmyslové zemi, alkoholismus se stal lidovou chorobou a počet sebevražd se řadí

mezi největší v Evropě. (=In Hungary, the number of deaths due to cardiac and vascular disease is more than any other industrialized country, alcoholism became a folk disease and suicide rates are almost the largest in Europe.)

9.1.6 Further underspecified group REST

The REST bridging relation is annotated in case when expressions are connected by a bridging relation which is not included in any of the groups above. This type is used for capturing potential candidates for a new group of bridging relations. If needed, this relation can be later relatively easily extracted from the annotated data.

Type REST includes the following groups of bridging relations:

1. The relation "location – resident" (by place names, e.g. *Praha (=Prague) - Pražák (*= lit. *resident of Prague)*), but also by common nous (e.g. *stát (=state) – obyvatelé (= population)).* Cf.

   a. Kwasniewski opakovaně zdůraznil, že z cesty zásadních proměn země nelze sejít: pravice a levice se budou přít se středem o tempo změn, ale o základní vzorec reforem není sporu. b. Co však je vážné, je nevelký zájem veřejnosti {bridging, type = REST to „země" in a.} *o věci veřejné. (=a. Kwasniewski has repeatedly emphasized that one cannot step aside from the way fundamental changes of the country … . b. However, the small public interest [lit. interest of the the public[in public affairs is important.)*

2. Relations between relatives (*matka* (= *mother) – syn (=son)*, etc.) Cf.

   a. Úzce navazuje na tradici podnikání svého rodu, především dědy. b. Od něj {coref_text to „děda"} *získal vnuk* {bridging, type = REST to „#PersPron"} *výtečné základy, ač sám vystudoval školu zaměřenou na dopravu. (=a. He … follows the traditions of his family business, especially of his Grandpa. b. He gave his grandson strong fundamentals, although he graduated from a traffic-orientated school.)*

3. The relation "author – his work", cf.

   Krásná, ale nesignovaná krajinka {no bridging relation} *neznámého malíře* {functor AUTH} *bude určitě hůře prodejná než slabý Slavíček.* (=*Beautiful, but unsigned landscape by unknown painter will be definitely not so popular as weak works of Slavíček.*)

4. The relation “event – argument” (*poslech* (*=listening) – posluchač (=listener))*

   a. V ČR bývají prostitutky zadržovány zpravidla jenom kvůli ověření totožnosti. b. Zákon o prostituci {bridging, type = REST to „prostitutky“} *se u nás teprve připravuje. (=a. In Czech Republic, prostitutes are usually controlled just for verification of identity. b. The law on prostitution is still being prepared.)*

5. The relation “object – typical instrument”

   a. Začal jsem, řekněme, jako provazochodec. b. Lidé chodili po zemi, já nějakých dvacet centimetrů nad ní. c. Klidně jsem mohl seskočit a dál dělat ve státním podniku, nic by se nestalo. d. Ale začal jsem lano {bridging, type = REST to „provazochodec“ in a.} *zvedat a seskočit už nebylo možné. (=a. I started, let's say, as a tightrope walker. ... d. But I started to lift a pole and it was not possible to jump off.)*

The participation on the text cohesion is considered to be important, so in ambiguous cases, those relations are annotated that are important for the text cohesion.

9.2 Limiting the number of bridging arrows

Given that the marking of bridging relations is very useful for information extraction, question answering and other NLP tasks, we decided to annotate them in PDT. However, this is a very complicated and time-consuming task, which up to now has not given high enough evaluation results. To make our annotation cleaner, we use the following conventions:

1. If ambiguous, prefer coreference over bridging relation (see Preference of coreference over bridging anaphora in 6)

2. Each node can only be an antecedent/anaphor for no more than one type of bridging relations. If more bridging types are possible, one is chosen according to the following scale:

SUBSET → PART → FUNCT → ANAF → CONTRAST → REST.

Cf. the following example where of two possible relations (SUBSET by *Podání intelektuálské (=the intellectual view)– tvář* (*=face*) and *Podání pragmaticky obchodní a ekonomické (=the pragmatic trade and economic view)– tvář (=face*)) and CONTRAST in the

pair of expressions *Podání intelektuálské* (=*the intellectual view*) and *Podání pragmaticky obchodní a ekonomické* (=*the pragmatic trade and economic view*)) we choose the first possibility.

> *a. Tento postoj má více tváří. b. Podání intelektuálské* {bridging, type = SET_SUB, to „tvář“ in a.} *pochází z pochybování o veškeré realitě včetně sebe samého a ústí v postmodernistický relativismus a neschopnost zaujmout pevný jednoznačný postoj. c. Podání pragmaticky obchodní a ekonomické* {bridging, type = SET_SUB, to „tvář“ in a.} *jednostranně preferuje krátkodobé, praktické potřeby, tedy potřeby zaměřené na současnost a bezprostřední budoucnost. (=a. This approach has several faces. b. The intellectual view comes from ... c. The pragmatic trade and economic view prefers short-term, practical needs, thus focusing on the present needs and immediate future.)*

Number of relations of the same type is not limited. It is quite common, that several arrows of types SUB_SET, PART_WHOLE begin in one node. Cf. the following example and Fig 13.

> *a. Nová striktní omezení vlády SR proti českým exportérům. b. Z téměř tří desítek smluv upravujících vztahy mezi oběma subjekty celního soustátí* {bridging, type = SUB_SET, to „SR“, bridging, type = SUB_SET, to „český“} *jsou okamžitě vypověditelné všechny... (=a. The new strict restrictions of Slovakia government against Czech exporters. b. Of the nearly three dozen agreements governing the relations between the two entities of the Customs unity, all of them should be immediately cancelled ...)*

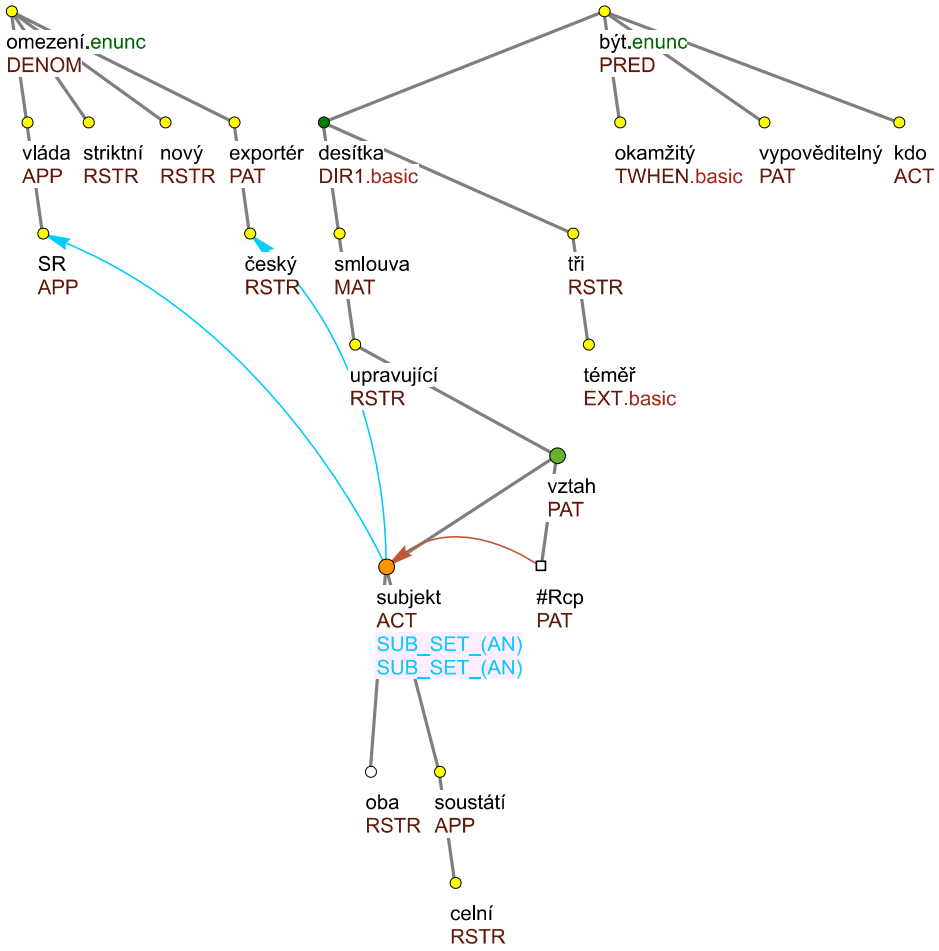

Fig. 13: Limiting the number of bridging arrows

3. Bridging relation is not marked if it can be extracted from the syntactical information annotated in the tectogrammatical layer in PDT. This concerns the following cases:

a. If a node has a tectogrammatical functor `APP`, `MAT`, `AUTH` or `PAT`, the bridging relation to its direct parent is NOT annotated as bridging relation.

Examples:

obyvatelka obce (= *resident of a village*) – possible REST, functor PAT

prezident Polska (= *president of Poland*) – possible FUNCT, functor PAT

dílo Wagnera (=*Wagner's work*) – possible REST, functor `AUTH`

> *Místopředseda* {no relation} *sněmovny* {no relation} *Jan Kasal uvedl pro LN, že s nesnázemi přijímá skutečnost, když někdo během rozehrané partie mění pravidla hry.* (=*The vice-president of the Upper Chamber Jan Kasal said LN that ...*)

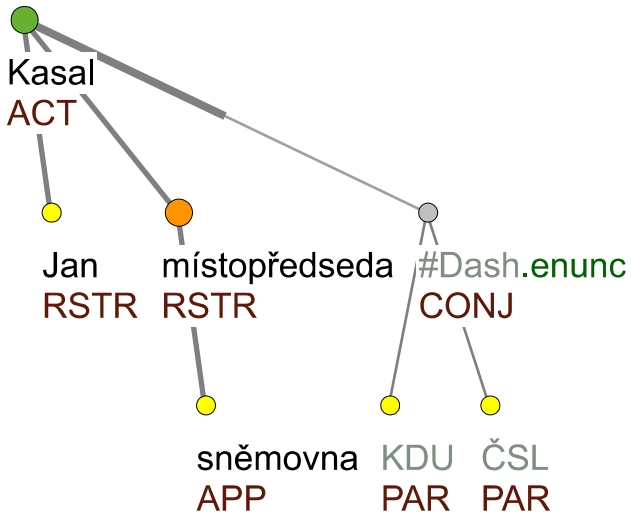

Fig. 14: Limiting the number of bridging arrows

This rule also holds for coordinative constructions and appositions, i.e. there would be no bridging relation of type FUNCT between *prezident (= president)* and *akciová společnost AC Sparta Praha – fotbal* (=*The stock AC Sparta Prague - football*) which are used in the context of the following sentence, because the dependent nodes of the comma have tectogrammatical functor APP. See also Fig. 15.

Pokusili jsme se telefonicky kontaktovat prezidenta AC Sparta Praha – fotbal, a. s., Petra Macha, ale jeho vyjádření, zda vyhoví asociaci, jsme nezískali. (=*We tried to phone the president of AC Sparta Praha - Football, Inc., Peter Mach, but …*)

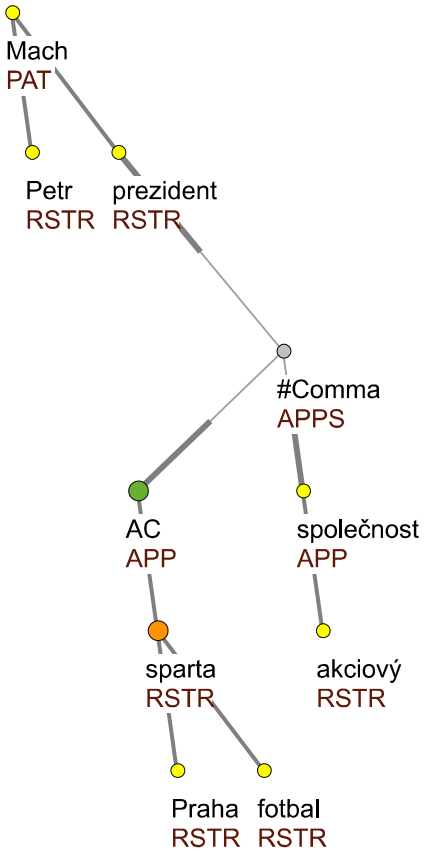

Fig. 15: Limiting the number of bridging arrows

b. Bridging relation of the type PART is NOT annotated if nodes are directly dependent one on the other, and dependent nodes have the tectogrammatical functor ACMP. Cf. *kaplička – daty narození, vlády a smrti* (=*chapel - dates of birth, death and the governing*) in the following sentence:

> *Na břehu Starnberského jezera u místa utonutí byla postavena kaplička s královými daty* {no bridging relation} *narození, vlády a smrti, s křížem* {no bridging relation} *a mramorovou pamětní deskou* {no bridging relation}. (=*On the shore of the lake, … there was built a chapel with the dates dates of birth, death and the governing of the king...*)

c. Bridging relation, type CONTRAST is NOT annotated if coherension-relevant discourse opposites are direct or indirect descendants of tectogrammatical nodes with functors ADVS (*adversative*) or CONFR *(confrontation)*. The meaning of contrast is already marked by these functors.

Examples:

Dočasný podnikatelův zisk bude anulován, ale.ADVS trvalý zisk {no bridging relation} *z jeho inovace zůstane zachován společnosti ve formě nižších cen nebo technicky dokonalejších výrobků. (=Temporary podnikatelův profit will be annulled, but.ADVS the permanent profit … will be retained in the form of lower prices or technically sophisticated products.)*

Letos by výstavba technického zařízení v sedmi lokalitách stála 120 milionů korun, ale.ADVS můžeme uvolnit jen 80 milionů {no bridging relation}*.(=This year, the construction of technical facilities in seven locations costed 120 million crowns, but.ADVS we can release only 80 million.)*

10 Problematic cases

10.1 Prepositional phrases

Not surprisingly, *in Prague* does not mean the same as *near Prague.* However, in tectogrammatical structure, prepositions are hidden in sub-functors and can be taken into account if annotating on tectogrammatical trees only by looking at these subfunctors. Although the semantic distinction between preposition phrases with the same head and different preposition is very important, we decided to ignore it in the annotation. So the following convention holds:

If two noun phrases are coreferential, we mark coreferential relation between them also in case when they are parts of prepositional phrases which are not coreferential.

Example:

a. Zatím se posunuje stále více za Prahu, čímž ztrácí na své účelnosti z hlediska dopravních spojení do jednotlivých částí města {coref_text, type = SPEC to „Praha“}. b. Na druhé straně by tu {coref_text, type = SPEC to „město“ in a.} asi mohlo být víc pozemků vhodných k podnikání. c. Po dálnici bychom se měli svézt z Prahy {coref_text, type = SPEC to „tu“ in b.} až do Českých Budějovic, v roce 1997 pravděpodobně projedou první vozidla po dálnici Praha – Plzeň, dokončena by měla být i dálnice D8 z Prahy do Ústí nad Labem. (=a. So far, people begin to move away from Prague, ...

various parts of the city . b. On the other hand, there could be more lands suitable for business there. c. Highways could take us from Prague up to Ceské Budejovice … .)

10.2 Specific syntactic construction of the type “faktory (= factors) - jeden z faktorů (= one of the factors)”

From the point of view of coreference annotation, the construction “*X – one of X*” is atypical. In a tectogrammatical tree, an expression *X* in *one of the X* has the DIR1 functor, so the coreference annotation inside the sentence cannot be omitted, as it is done for similar cases in 9.2. For this reason, the decision of annotating such constructions is the following:

1. NPs with the governing node *X* are linked by SPEC or GEN coreference according to their reference type.

2. The word *one* is linked to the second X be the means of the bridging relation, type SET_SUB, as shown on Fig. 16.

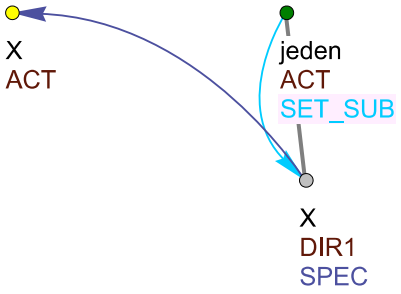

Fig. 16: "factors" - "one of the factors"

10.3 Specific syntactic constructions of the type “zaměstnanci (= employees) - každý ze zaměstnanců (= *lit.* each of the employees *meaning* all of the employeers)”

The construction *“X – each of X”* seem similar to that in section 10.2, but it has different referential value. Here, *each of* is coreferential with *X* and it is in the position of a syntactic noun. For this reason, this construction is annotated as a whole, similarly as for the cases where quantifier is in the position of a semantic adjective. Coreference is marked with the dependent node X.

Example:

a. Portmonky měla v referátu rodina Dunkových. b. Všichni to věděli a všichni na ně byli krátcí, protože Dunků bylo moc a každý z nich měl svá lidská a občanská práva. c. Z pálení ukradených peněženek se kolem jejich chalupy linul penetrantní čmoud. (=a. The Dunk family had the case of wallets. b. Everyone knew that and did not say a lot, because there was a lot of Dunks there and each of them had his human and civil rights.)

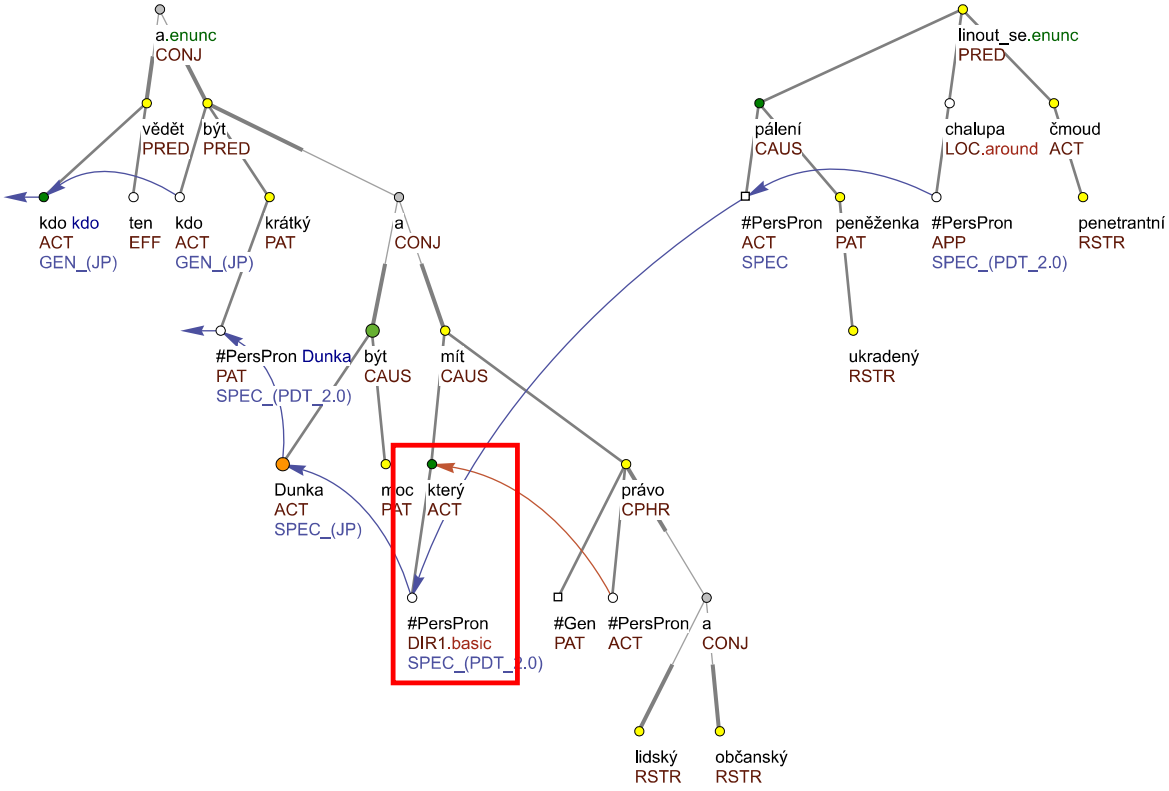

Fig. 17: Specific syntactic construction of the type "zaměstnanci (= employees) - každý ze zaměstnanců (= lit. all of the employees)"

11 Annotation in TrEd

11.1 The Tree Editor TrEd

The primary format of PDT 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 PDT 2.0 has been enriched to support the annotation of the extended textual coreference (which has – unlike the originally annotated textual coreference – a type) and the bridging anaphora (which has not been annotated before and also has a type). Technically, various kinds of non-dependency relations between nodes in PDT 2.0 use dedicated referring attributes that contain unique identifiers of the nodes they refer to.

11.2 The Annotation Tool

We employ two ways of helping the annotators in their tedious task. First, we pre-annotate the data with highly probable coreference relations. The annotators check these links and can remove them if they are wrong. This approach has proved to be faster than letting the annotators annotate the data from scratch. Second, we have implemented several supporting features into the annotation tool (the TrEd extension) that help during the annotation process.

11.2.1 The pre-annotation

We use a list of pairs of words that with a high probability form a coreferential pair in texts. Most of the pairs in the list consist of a noun and a derived adjective, which are different in Czech, e.g. *Praha (= Prague) – pražský (*=lit. *Prague)*, like in the sentence: *He arrived in Prague and found the Prague atmosphere quite casual*). The rest of the list is formed by pairs consisting of an abbreviation and its one-word expansion, e.g. *ČR – Česko* (similarly in English: *USA – States*). The whole list consists of more than 6 thousand pairs obtained automatically from the morphological synthesizer for Czech, manually checked and slightly extended.

This pre-annotation concerns only the cases of textual coreference. Bridging relations are not pre-annotated.

11.2.2 The Annotation

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

11.2.2.1 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.

11.2.2.2 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. 18 (krok 1 = step 1, krok 2 = step 2):

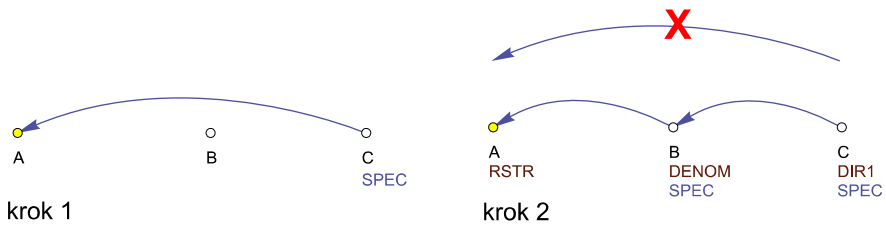

Fig. 18: Finding the nearest antecedent

11.2.2.3 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. 19:

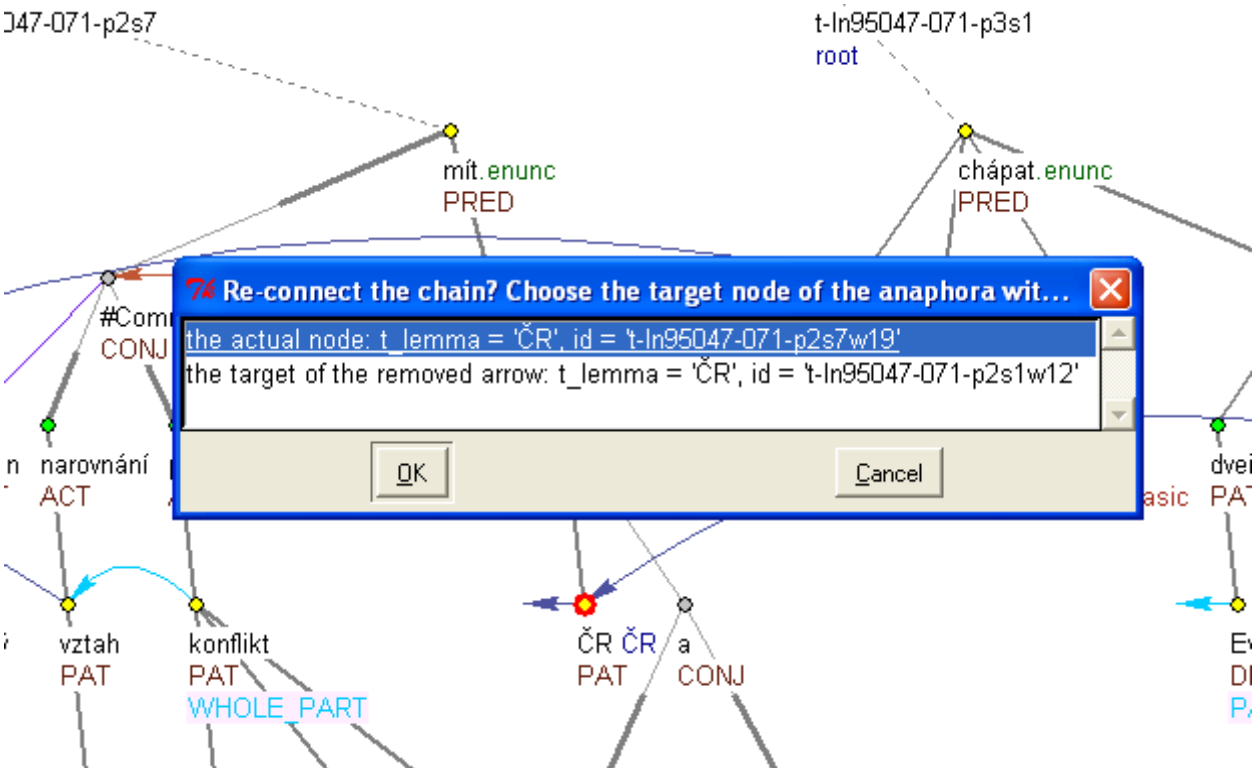

Fig. 19: Preserving the coreferential chain

11.2.2.4 Text highlighting

The annotation of extended textual coreference and bridging relations is performed on the tectogrammatical layer of PDT. However, the annotators prefer to work on the surface form of the text, 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. 20:

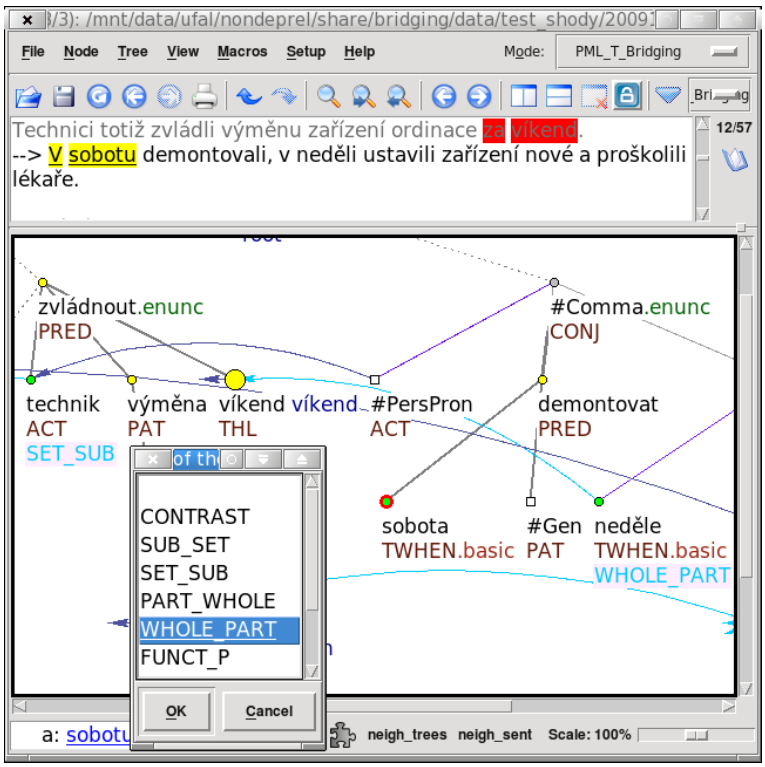

Fig. 20: Text highlighting

11.2.2.5 Bridging long coreferential chains

When two coreferential chains are also in a bridging relation, the bridging relation is marked only once, by the first appearance of the second coreferential chain. This principle is controlled automatically, as a post-annotation correction. Bridging relations ANAF and CONTRAST are excluded from this principle.

Schematically, for the nodes *Alžír* (=*Algeria*), *Alžírský (=Algerian) – Alžířan (=resident of Algeria)*, this rule is exemplified on Fig. 21:

Fig. 21: Bridging long coreferential chains

11.2.3 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.

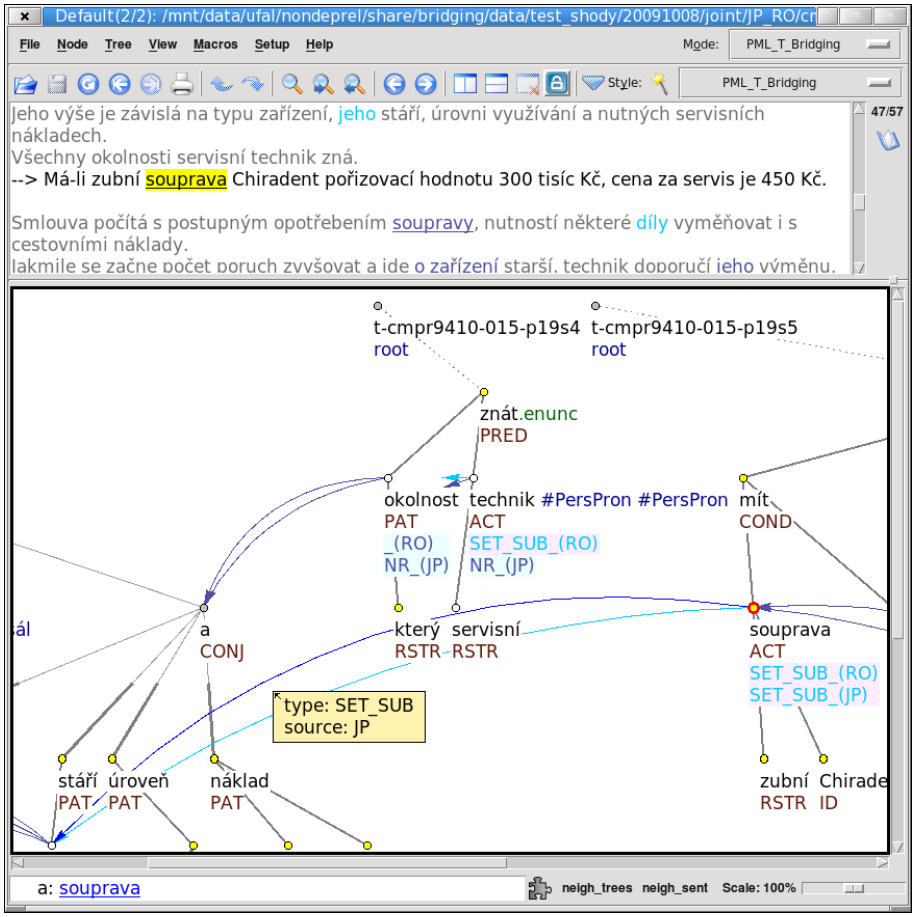

Fig. 22: Comparing different annotations

References

CLARK, Herbert H. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, June 10–13, Cambridge, Massachusetts, 1977.

CHIARCOS, Christian; KRASAVINA, Olga. PoCoS – Potsdam Coreference Scheme. In *Proceedings of ACL-2007 Linguistic Annotation Workshop*. Praha, 2007, s. 156–163.

KUČOVÁ, Lucie; KOLÁŘOVÁ, Veronika; ŽABOKRTSKÝ, Zdeněk; PAJAS, Petr, ČULO, Oliver. *Anotování koreference v Pražském závislostním korpusu*. Praha: UFAL/CKL MFF UK, 51, Technická zpráva-2003-19, 2003.

KUČOVÁ, Lucie; HAJIČOVÁ, Eva. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of 5th Discourse Anaphora and Anaphor Resolution Colloquium*. Edicoes Colibri, 2004.

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.

MLADOVÁ, Lucie; ZIKÁNOVÁ, Šárka; HAJIČOVÁ, Eva. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*. Marrakech, Morokko, 2008.

MÜLLER, Christoph; STUBE, Michael. Annotating anaphoric and bridging relations with MMAX. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, 2001, s. 90–95.

PADUČEVA, Elena V. *Vyskazyvanije i jego sootnesennost' s dejstviteľnostju*. Moskva: Nauka, 1985.

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.

POESIO, Massimo. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL*. Boston, duben, 2004.

RECASENS, Marta; MARTÍ, Antònia. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. In *Language Resources and Evaluation*. 2010.

VIEIRA, Renata; POESIO, Massimo. An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics*, roč. 26, č. 4, 2000, s. 539–593.