The Prague Bulletin of Mathematical Linguistics 78, 2002

PDT: Two Steps in Tectogrammatical Annotation with respect to some Issues of Deletion

Veronika Řezníčková

Abstract

The annotation of the Prague Dependency Treebank is realized in two sub-collections which differ in the subtlety of annotation (the large collection and the model collection). In the present paper, we focus on deletions of complementations of verbs, postverbal nouns and adjectives, from the point of view of the annotators of the model collection. We inquire into the issues of deletions of participants of verbs with respect to the coreferential relations between the restored node and its antecedent, and we introduce the new type of deletion where the deleted node has no concrete antecedent (we call these restored nodes vague, unspecified anaphoric elements). We also specify the differences between the newly introduced lemma *Unsp(ecified)* and the lemma *Gen(eral)*, used for deletions of General Participants. After analysing the process of nominalization, we extend the description of the particular types of deletions also to the postverbal nouns denoting action.

1. Three-layer system of tags

The annotation of the Prague Dependency Treebank (PDT in the sequel) is basically conceived of in accordance with the theoretical assumptions of the Functional Generative Description (FGD in the sequel, see Sgall, Hajičová, and Panevová, 1986; Hajičová, 1993). The present paper deals with the current phase of annotation of PDT at the so-called tectogrammatical level (which captures underlying syntactic structures of sentences). This level of annotation, in contrast to the two preceding phases of annotation, namely the morphemic and the so-called analytical levels (for a description of the annotation scheme of PDT see e.g. Hajič et al., 2001, Hajičová et al., 2001), is realized in two steps (in two sub-collections) which differ in the subtlety of annotation. The first, basic step of annotation is represented by the so-called basic or large collection (LC; today it contains about 25,000 sentences). The second step of annotation, the so-called model collection (MC), should provide full, detailed information about the underlying structure of the sentence. Due to the fact that this way of annotation is more detailed, therefore more elaborate, the model collection contains only about 400 sentences today.

2. General principles of annotation at the tectogrammatical level

Tectogrammatical tree structures (TGTSs) are based on dependency syntax; they have the shape of a dependency tree with the verb as the root of the tree and its daughter nodes representing nodes depending on the governor (on each layer of the tree). The two dimensions of the tree represent the syntactic structure of the sentence (the vertical dimension) and the topic-focus articulation of the sentence, based on the underlying word order (the horizontal dimension).

The tagging at the tectogrammatical level can be described by the following principles:

(a) a single node of a TGTS may be a representation of more than one word; only autosemantic words have a node of their own, while the correlates of functional words (auxiliaries, prepositions etc.) are attached to the autosemantic words to which they belong (auxiliary verbs and subordinating conjunctions to the verbs, prepositions to nouns, etc.);

(b) in the cases of deletion in the surface shape of the sentence, nodes are introduced into the tectogrammatical tree to 'recover' a deleted word;

(c) no non-projective structures are admitted at the tectogrammatical level (projectivity is a counterpart to continuity of constituents; non-projectivity appearing in the surface shape of some sentences is supposed to be handled by movement rules between the tectogrammatical tree and the morphemic string);

(d) not only the direction of the dependence on the governing node (dependence to the left, dependence to the right) is taken into account, but also sister nodes are ordered (from left to right); their order reflects the scale of communicative dynamism which is relevant for the description of topic-focus articulation.

3. Complex tags

Each label of a node consists of the following parts:

- the lexical value proper of the word (represented in a preliminary way just with the usual graphemic form of the word, the **lemma**),

- the values of the **morphological grammatemes** (corresponding primarily to the values of morphological categories such as modality, tense, aspect with verbs, gender and number with nouns, degree of comparison with adjectives),

- the values of the attribute **functor**, corresponding to (underlying) syntactic functions (Actor, Objective, Means, Locative, etc.);

- the values of the attribute **syntactic grammateme** (i.e. a more subtle classification of functors), accompanying some of the functors according to a more subtle (semantic) differentiation of these syntactic relations that is rendered on the surface first of all by prepositions and cases of nouns; this concerns the functors with participants (arguments) ACT (Actor) and PAT (Patient); e.g. *Není peněz* '(There) is no money' ACT.GNEG; *Vody ubývá* 'Water (Genitive) is running low' ACT.GPART; then it concerns the functors with the meaning of location LOC, DIR-1, DIR-2 and DIR-3 (corresponding to the questions 'where?', 'from where?', 'through which place?' and 'where to?', respectively); thus e.g. LOC (expressed in Czech by several prepositions which combine either with the locative (Loc) or with the instrumental (Instr) case of the noun) is subcategorized into *na*+Loc ('on': *na stole* 'on the table'), *v*+Loc ('in'), *u*+Loc ('by'), *nad*+Instr. ('above'), *pod*+Instr ('under'), *za*+Instr ('behind'), *mezi*.1+Instr ('among'), *mezi*.2+Instr ('between'), etc. As for functors having a temporal meaning, a similar subcategorization is established with the functor TWHEN (with the grammatemes AFT 'after', BEF 'before', NIL 'on Monday', 'next year'). A positive or negative grammateme is attached to ACMP ('with' vs. 'without'), REG ('with regard' vs. 'without regard') and BEN ('for' vs. 'against');

- the values of a special attribute capture the basic information about the **topic-focus articulation** (TFA) of the sentence;

- the values of the attribute **del(etion)** are specified for nodes restored in the tectogrammatical tree structure and correspond to the different character of the deleted item, often depending on its antecedent. When the antecedent of the node is an expanded head node, then the restored node gets the index ELEX. If this is not so, then the restored node gets the index ELID;

- the values of the attributes **coref(erence)**, **cornum** (number of antecedent), **corsnt** (coreference in the sentence) and **antec(edent)** (functor of antecedent with grammatical coreference) which reflect the linking of sentences to each other and to the context of situation. The value of the attribute COREF is the lexical value of the antecedent of the given anaphoric node (this node itself may be present on the surface or deleted). The value of the attribute CORNUM is equal to the serial number of the antecedent of the given node (to avoid uncertainty in case of two occurrences of the same word in the sentence). The attribute CORSNT indicates whether the antecedent is in the same sentence (it gets the value NIL)

or in the preceding context (it gets the value PREVi, e.g. PREV1 if the antecedent is in the immediately preceding sentence, PREV2 if the antecedent is in the second sentence preceding the given sentence, PREV3...)

- some other indices for **phrasemes**, **direct speech** etc.

4. Two sub-collections: Large collection and Model collection

The two sub-collections of the tectogrammatical level of annotation (namely the large collection (LC) and the model collection (MC)) differ especially in the entirety of filling in the values of the attributes mentioned above. After the first automatic procedure which allows for a transduction of the analytic trees to the tectogrammatical ones the main task of annotators of LC is to revise the syntactic structure of the given sentence, to restore the deleted nodes and to label all nodes in the tree with functors. The main task of annotators of MC is to restore remaining elided nodes, to fill in the values of the attributes mentioned above which have not been derived by automatic transduction from the analytic level and to capture the transposed use of forms (historical present tense and the present *pro futuro*, singular validity of *pluralia tantum*, etc.).

5. Deletions of complementations of verbs, postverbal nouns and adjectives

One of the most difficult issues within corpora annotation on an underlying syntactic level is the restoration of nodes omitted in the surface shape of the sentence, but present on the level of tectogrammatical structure (TGTS). We would like to illustrate the complexity of the task of the annotation of the "model" collection on some of the issues concerning the restoration of nodes for semantically obligatory complementations (valency slots) of verbs, postverbal nouns and adjectives, and on the issues concerning coreference relations of the restored nodes to their antecedent.

5.1. Deletions of complementations of verbs

5.1.1. Classification of deletions

We have already mentioned that in the cases of deletion in the surface shape of the sentence, nodes are introduced into the tectogrammatical tree to 'recover' a deleted word. Some of the papers mentioned in the list of references deal with the issues of deletions, i.e. with the general principles of the reconstruction of deletions. According to these papers the following types of deletions should be recognized (see esp. Hajičová and Ceplová, 2000; Hajičová and Sgall, 2000):

(i) deletions licensed by the grammatical properties of sentence elements or sentence structure (grammatical identification of the deleted item).

Within group (i), two situations may obtain:

(a) only the syntactic position itself in the sentence structure is predetermined, but its lexical setting is „free" (esp. the zero form of a subject pronoun given by the pro-drop character of Czech, e.g. *Předseda vlády řekl, že předloží návrh na změnu volebního systému* 'The Prime-minister said that Ø will submit a proposal on the change of the electoral system' - The "dropped" subject of the verb předloží "will submit" may refer to the Prime-minister, to the Government, or to somebody else identifiable on the basis of the context)

(b) both the position and its „filler" are predetermined (as with verbs of "control", e.g. *Předseda vlády slíbil předložit návrh na změnu volebního systému* 'The Prime-minister promised to submit a

proposal on the change of the electoral system' – The identification of the underlying subject of the infinitive is "controlled" by the Actor of the main verb, in this example it is "the Prime-minister")

A specific type of deletion is that of the so-called General Participant (see Daneš 1971, Panevová 1973). Its "filler" may be paraphrased as "those who in general are competent to do it" or "that what in general is used to be done in such a case" and so on, e.g. *Náš chlapec už čte* 'Our boy already reads'. – What does he read? That what in general is used to be read.

(ii) deletions possible only if the preceding context (be it co-text or context of situation) exhibits certain specific properties (contextual identification of the deleted item); this type of deletion may be called textual ("occasional") deletion (e.g. *Potkal jsi Jirku? Potkal.* 'Have you met Jirka? (I-) Met (him).')

From the other point of view, considering the possibility / impossibility of restoration of the respective node at the surface layer of the sentence, we can differentiate three possibilities within group (i):

(ia) It is possible to restore the deleted node.

(iba) In constructions with the obligatory "control" it is "forbidden" to restore the deleted node from clear grammatical reasons: it is simply impossible to restore it. Also in passive reflexive constructions with the general Actor a word expressing the Actor is excluded.

(ibb) In sentences with general Patient it is also "forbidden" to restore the deleted node. F. Daneš states that in sentences with this type of participant the object is not expressed and it is not possible to restore it (e.g. *Tenhle pes nekouše* 'This dog does not bite'; Daneš, 1971, p. 133). The syntactic position for the Patient is here, of course, open; however, any filler (including various paraphrases of the notion of generalization) changes the meaning. Paraphrases like *Tenhle pes nikoho nekouše* 'This dog does not bite anybody', *Náš chlapec už čte všechno, co je možné číst* 'Our boy already reads everything that is possible to be read' are neither common, nor equivalent to the empty surface form.

5.1.2. Possible lemmas of restored nodes with respect to the type of coreference

All types of deletions mentioned above are accompanied by a specific type of coreferential relations between the restored node and its antecedent – they were paraphrased by such expressions like "lexical setting" or "filler" there. This coreferential relation may be again grammatical or textual (cf. Hajičová, Panevová, and Sgall, 2000). Annotators have to label restored nodes with special lemmas (*Cor* "control" for the cases of grammatical coreference, *Gen(eral)* for General Participants , *já* 'I', *ty* 'you', *on* 'he' etc. for the cases of textual coreference) according to the type of deletion and the type of coreference, and to specify the values of the attributes (i.e. attributes COREF, ANTEC, CORNUM and CORSNT, cf. above).

Thus the coreferential relation between the restored node and its antecedent may be:

a) **Grammatical**, as that between an argument of the verb of control (functioning as the controller) and the subject of the embedded verb (functioning as the controllee) (the group (ib) above), e.g. in the TGTS for the sentence *Podnik hodlá zvýšit výrobu* 'The company intends to increase the production' a node is restored depending on the verb *zvýšit* 'increase' with the lexical label *Cor* and the functor ACT; the attributes of the coreferential relations get the relevant values.

b) **Textual**, as with the zero form of a subject pronoun given by the pro-drop character of Czech (the group (ia) above) or with the deletion of the respective node in the surface shape of the sentence which

is conditioned by the preceding context rather than by some grammatically determined conditions (the group (ii))[1].

c) **Transition from textual coreference to coreferential relations without a concrete antecedent**

It seems to be necessary to distinguish also deletions which do not refer to any concrete antecedent; thus the attribute COREF with the restored node gets the value NA (=non-applicable). This is not only the case of General Participant (belonging to the group (i)), but it concerns also restored nodes with pronominal lemmas, especially with "deconcretized" 1st and 3rd person plural (cf. also Panevová 1998, e.g. *V češtině máme sedm pádů* 'In Czech we-have:1st pl seven cases', *Tady dobře vaří* 'Here well they-cook: 3rd pl'). Analysing the sentences from PDT, we find also sentences where the restored node refers to the "contents" of the preceding text rather than to some particular element.

5.1.3. Introduction of a new lemma *Unsp*

While discussing the issues of "deconcretized" participants, J. Panevová postulated an **unspecified, vague** anaphoric element *oni* 'they' for labelling this type of nodes restored on the underlying level (see Panevová, 1998). The vagueness of this element is reflected by the introduction of the special lemma *Unsp(ecified)* instead of clear anaphoric character of the lemma *on* 'he' (see also Hajičová et al., 2001)). According to Marková and Panevová (in press) the lemma *Unsp* is used when the node refers to some not exactly specified group of people / objects. Lemma *Unsp* serves only for those cases of deconcretized Actors which do not include the speaker (therefore only for those constructions where the form of the verb is in agreement with the zero subject form "3rd person plural, animate"). Marková and Panevová mention also one of the distinctions between two special types of deletions, namely between the lemma *Gen* and the lemma *Unsp*: The possibility of a (at least) probable delimitation of the referent distinguishes the lemma *Unsp* from the lemma *Gen* which is used for a node referring to all Actors / Patients typical for the respective situation.

The specification of criteria for distinction between the lemma *Gen* and the lemma *Unsp* seems to be a very delicate issue. We will inquire into these criteria analysing individual participants, namely Actor, Addressee and Patient.

For the differentiation of the compared lemmas with all these participants two properties, serving also as the different points of view for classification of deletions, are crucial:

(i) typical surface realization (cf. section 5.1.1 above);

(ii) possibility / impossibility of delimitation of a referent (cf. section 5.1.2 above).

5.1.3.1. *Unsp* with Actor

According to Marková and Panevová also the exclusion of the speaker can differentiate between the two investigated types of deletion when Actor is concerned. Table 1 on the next page illustrates possible criteria for the distinction of the lemma *Gen* and the lemma *Unsp.*

Deletions with General Participants:

(1) *Tato potvrzení se vydávají Gen.*ACT *Gen.*ADDR *na počkání*

(1') 'These receipts Refl issue-3rd pl upon waiting'

[1] We do not capture coreferential relations in cases of deletions with pronominal lemmas for the 1st and the 2nd person (the lemmas *já* 'I', *ty* 'you', *my* 'we', *vy* 'you'), even not in the sentences where the referent can be delimited from the context, e.g. *Na hřiště přišel i Honza.* 'Also Honza has come to the playground.' *„Zkusíš* (*ty* 'you' ELID, i.e. Honza) *to taky?", zeptal se Petr a ukázal na Honzu.* 'Will (you ELID) try it too?', Petr asked and pointed at Honza. *„Jen to přeskočíš* (*ty* 'you' ELID, i.e. Honza) *a podlezeš."* '(You) will just jump over it and crawl under it.'.

Lemma of the node	Exclusion of the speaker	Typical surface realization	Delimitation of the referent
Gen	It does not say anything	Reflexive passive construction	All Actors typical for the respective situation
Unsp	Yes	The verb is in agreement with the zero subject form "3rd person plural, animate"	A not exactly delimited group of people, but we can estimate the possible referent from the context

Table 1: Differences between the lemma *Gen* and the lemma *Unsp* with ACT

Deletions with unspecified elements:

(2) *Ukradli Unsp.*ACT *nám auto*

(2') 'They-stole the car from us'

(3) *Vypnuli Unsp.*ACT *proud*

(3') 'They-cut off electricity'

For constructions with unspecified Actors it is also typical, but not necessary, to use some adverbial with the meaning of location delimiting the group of people from which we can assume the possible referent, e.g.:

(4) *Na poště.*LOC *zavírají Unsp.*ACT *v šest hodin odpoledne*

(4') 'At the post-office.LOC they-close at six o'clock p.m.'

(5) *Tady.*LOC *dobře vaří Unsp.*ACT

(5') 'Here.LOC well they-cook'

5.1.3.2. *Unsp* with Addressee

Considering the Addressee indicates very often some man or group of people it is obvious that some adverbial with the meaning of location delimitating the group of people from which we can predetermine the possible referent is characteristic also for sentences where deletions of Addressee can be restored with the newly introduced lemma *Unsp*, e.g.:

(6) *Doma.*LOC *slíbil Unsp.*ADDR*, že přijde brzy, ale kamarádům řekl něco jiného.*

(6') 'At home.LOC he promised Ø to come soon, but he told his friends something else'

(7) *Celý den o tom jednal Unsp.*ADDR *ve vládě.*LOC.

(7') 'The whole day he negotiated Ø about that in the cabinet.LOC'

In contrast with the same type of deletion with Actor, the zero position can be treated as typical surface realization of deletion of vague, unspecified Addressee. Of course, the criterion characteristic just for Actor (exclusion of the speaker) is missing in the table describing the criteria for distinction between the lemma *Unsp* and the lemma *Gen* (cf. also example (1) above) with Addressee (see Table 2):

Lemma of the node	Typical surface realization	Delimitation of the referent
Gen	Zero	All Addressees typical for the respective situation
Unsp	Zero	A not exactly delimited group of people, but we can estimate the possible referent from the context

Table 2: Differences between the lemma *Gen* and the lemma *Unsp* with ADDR

5.1.3.3. *Unsp* with Patient

Two investigated types of deletion with Patient seem to be distinguishable only with difficulty. The zero object position is the typical surface realization with both of them. However, the possibility / impossibility of delimitation of an referent (sometimes just from the context, but sometimes also by some adverbial with the meaning of location identifying the group of possible referents) helps to differentiate the lemma *Gen* from the lemma *Unsp* also with Patient (see Table 3):

Lemma of the node	Typical surface realization	Delimitation of the referent
Gen	Zero	All Patients typical for the respective situation
Unsp	Zero	A not exactly delimited group of Patients, but we can estimate the possible referent from the context

Table 3: Differences between the lemma *Gen* and the lemma *Unsp* with PAT

Deletions with unspecified elements:

(8) *Ptáme se na vrcholnou sezónu: „Kdy je nejlepší přijet na safari?“.*

(9) *„Od poloviny prosince do poloviny února, vybrali jste si Unsp*.PAT *dobře“, chválí Unsp*.PAT *John.*

(8’) We ask about the peak season: “What is the best time to come to safari?”

(9’) “From the middle of December to the middle of February, you have chosen Ø well”, commends Ø John.

Štícha (1987) describes several types of objects (referents) which can be referred to within constructions with an unexpressed object. Some of them can be also classified as deletions with the lemma *Unsp*:

(10) *Jakmile měla trochu času, už gruntovala, vynášela Unsp*.PAT, *přestavovala Unsp*.PAT...

(10’) ‘In her spare time, she was always tidying up Ø, carrying out Ø, rearranging Ø...’

(11) *Když se točila u plotny*.LOC*, míchala Unsp*.PAT*, přisypávala Unsp*.PAT*, přilévala Unsp*.PAT...

(11’) ‘When she was spinning around the kitchen range.LOC, she was mixing Ø, adding Ø, pouring Ø...’

According to Štícha, there are also some verbs which are commonly used without its object; the intransitive usage is very usual (fixed) with them. Their object can be almost explicitly determined, even

though it is not specified in the preceding context, e.g.: *smeknout* 'take off', *zaparkovat* 'park', *utrácet* 'overspend', *zapálit si* 'light (a cigarette) for oneself', *zavěsit* 'hang up'. The zero object position can refer to a single specific object, e.g. *sluchátko* 'receiver' with *zavěsit* 'hang up', or to a limited group of objects, e.g. *pokrývka hlavy* 'headgear' with *smeknout* 'take off', *vozidlo* 'vehicle' with *zaparkovat* 'park'.

There are three possibilities of capturing this type of deletion in FGD (PDT):

1. The verb will have two meanings in the lexicon: (i) the transitive verb with the possibility of using of the object; (ii) the intransitive verb;

2. The construction may be considered as a textual deletion, therefore the node will be restored with the pronominal lemma and the attribute COREF will be filled in by the lemma of the respective word, although this referent is not expressed in the preceding context;

3. The construction may be captured by the new lemma *Unsp* and in this way we will indicate the idiomatic intransitive verbs.

The selection of one of these possibilities is still an open question.

5.2. Deletions of complementations of deverbative nouns

The restoration of deletions includes not only the restoration of all obligatory participants and obligatory free modifications of verbs deleted for various reasons at the surface shape of the sentence, but also the restoration of obligatory members of valency frames of postverbal nouns and adjectives[2]. Description of particular types of deletion with postverbal nouns and adjectives is a very specific task and to be able to manage it, we must, first of all, understand the complexity of the process of nominalization. In this part of the paper, we focus on deletions of complementations of deverbative nouns denoting action.

5.2.1. The process of nominalization

By nominalizations we understand:

a) Nouns derived from verbs by productive means (e.g. *rozhodnutí* 'decision making', *obžalování* 'accusing' or nouns derived from verbs by non-productive means or by the zero suffix (e.g. *rada* 'advise', *slib* 'promise');

b) Nouns derived from a predicative adjective (e.g. *on je schopen udělat* 'he is able to do sth' → *jeho schopnost napsat knihu* 'his ability to write a book', *on je povinen udělat* 'he is obliged / required to do sth' → *jeho povinnost vydat majetek* 'his duty / obligation to release possession');

c) Nouns which primarily are a part of an analytical predicate (e.g. *Petr má šanci vyhrát* 'Peter has a chance to win' → *Petrova šance vyhrát* 'Peter's chance to win', *Petr má právo odvolat se* 'Peter has a right to appeal' → *Petrovo právo odvolat se* 'Peter's right to appeal');

[2] All obligatory dependents of verbs deleted at the surface shape of the sentence are restored in LC. As for nouns, the situation is more difficult. Only obligatory dependents of nouns derived from verbs by productive means (with the suffix *-ní, -tí*; so-called "verbální (slovesná) substantiva", e.g. *létat – létání* - 'to fly – flying'; *pokrýt - pokrytí* - 'to cover - covering') are restored in LC. Nouns derived from verbs by non-productive means (so-called "podstatná jména dějová", e.g. *létat - let* - 'to fly - flight') get their dependents only in MC. Also the deleted obligatory members of valency frames of postvebal adjectives are restored only in MC.

d) Deverbative adjectives (e.g. *dívka usiluje studovat* 'the girl intends to study' → *dívka usilující studovat* 'a girl intending to study', *dopis, který napsal Petr* 'letter that Petr wrote' → *dopis napsaný Petrem* 'a letter written by Peter').

5.2.1.1. Types of derivation during the word-formation process

While describing valency frames of nouns derived from verbs, Panevová (2000) differentiates, according to J. Kuryłowicz, two basic types of word-formative process: **syntactic derivation** and **lexical derivation**. While in syntactic derivation only syntactic function of derived word changes, in lexical derivation not only syntactic function, but also the lexical meaning of the derived word changes (see Kuryłowicz, 1936).

a) Syntactic derivation can be exemplified by the following nouns derived from verbs:

Čtenář si ověřuje danou informaci v jiných zdrojích → Ověřování dané informace čtenářem v jiných zdrojích 'The reader verifies the given information in other sources → Verification of the given information by the reader in other sources';

Odborníci stavějí katedrálu → stavění chrámu odborníky / stavba chrámu odborníky 'Experts build the cathedral → building of the cathedral by experts';

b) Nouns derived by lexical derivation are represented especially by actor names (e.g. *učit → učitel* 'to teach → teacher'), names denoting the place of action (e.g. *umývat → umývárna* 'to wash → washroom') and names denoting a tool (e.g. *ořezávat → ořezávátko* 'to sharpen a pencil → a pencil sharpener').

5.2.1.2. Incorporation of participants or free modifications

It is typical for nouns derived from verbs by lexical derivation that one of the inner participants or free modifications of the source verb is incorporated into the meaning of the noun itself, as the Actor with actor names, the place of action with names denoting the place and the tool with names denoting a tool (cf. Panevová, 2000). When incorporation of inner participants is concerned (not only the incorporation of Actor, but also that of Patient (e.g. *dárek mé sestře*.ADDR 'the gift to my sister', *výrobek firmy*.ACT 'a product of the firm') or Effect (names of artefacts which can be understood as a result of an activity, e.g. *hra o životě*.PAT 'the play about life'), the meaning of the noun itself occupies one of the valency slots of the source verb.

5.2.1.3. The process of substantivization

The incorporation of participants can be accompanied also by the process of substantivization. While the noun with an incorporated participant can keep other participants from the valency frame of the source verb (cf. the examples in the section above) and loses only the incorporated one, during the process of substantivization the noun loses gradually all its participants (the syntactic function of the dependent expressed by a possessive pronoun / adjective changes from Actor (ACT) to Appurtenance (APP)).

So we can differ several stages with the same noun:

1) noun derived from verb by syntactic derivation: noun denoting action (cf. examples (12), (14), and (16))

2) noun derived from verb by lexical derivation:

2a) noun with an incorporated participant (cf. examples (15), (17), and (18))

2b) noun after the process of substantivization (cf. examples (13), (15), and (18)).

(12) *Psaní dopisu.*PAT *Petrem.*ACT *trvalo asi 3 hodiny*

(12') 'Writing a letter.PAT by Peter.ACT took him about three hours'

(13) *Petrovo.*APP *psaní leželo na stole*

(13') 'Peter's.APP letter lay on the table'

(14) *Petrova.*ACT *výhra milionu.*PAT *šokovala celou rodinu*

(14') 'Winning of one million.PAT crowns by Peter.ACT shocked the whole family'

(15) *Petrova.*APP *výhra činila milion korun*

(15') 'Peter's.APP win was one million crowns'

(16) *Naše*.ACT *výplata dividend*.PAT *klientům*.ADDR

(16') 'Our.ACT payment of dividends.PAT to clients.ADDR'

(17) *Nemá na výplaty zaměstnancům*.ADDR

(17') lit. '(He) has not for payment to employees.ADDR' ('He has no money for employee wages')

(18) *To je moje*.APP *první výplata!*

(18') 'This is my.APP first wage-packet!'

These stages correspond to the particular meanings of the lexeme and express the transition of this lexical unit from a noun denoting action to a noun denoting substance.

5.2.1.4. "Overload" of the valency frame of a nominalized (condensed) verbal structure

Panevová (2000) presumes the "middle" position of particular types of deverbal nouns on the virtual axis: nouns derived from verbs by syntactic derivation (nouns denoting action) – nouns derived from verbs by lexical derivation (nouns denoting substance). Since the nominalized structure may get "overloaded", the partial process of substantivization, i.e. the weakening of "action usage", may be accompanied by a reduction of the number of slots in the valency frames in comparison with the frames assigned to the source verbs, although the respective noun still denotes action, e.g.:

(19) (*On*.ACT) *dokázal příteli.*ADDR *svou nevinu.*PAT → *jeho.*ACT *důkaz neviny.*PAT *?příteli.*ADDR

(19') 'He.ACT proved his innocence.PAT to (his) friend.ADDR → his.ACT proof of innocence.PAT ?to (his) friend.ADDR'

5.2.2. Deletions of complementations of deverbative nouns with respect to the type of coreference

It is already a very well-known fact that any of the participants of deverbative nouns can be omitted on the surface layer. It is typical for Czech sentences in newspaper articles that the writer uses a lot of deverbative nouns denoting action as the tool of condensation of information involved in the text. To exemplify that, let us present one paragraph from PDT:

(20) *Rozšíření na východ je podmínkou stability* (title)

(21) *Dánský ministr zahraničí pro Lidové noviny* (subtitle)

(22) *Jedním z hlavních témat řady rozhovorů dánského ministra zahraničí Nielse Helvega Petersena v Praze byla kromě bilaterálních vztahů i současná situace v Evropské unii.*

(23) *Dánská cesta k přijetí Maastrichtské smlouvy nebyla vůbec snadná: první celonárodní hlasování smlouvu odmítlo.*

(24) *Bylo zapotřebí řady měsíců přesvědčování, aby se v opakovaném hlasování Dánové nakonec těsnou většinou vyjádřili pro Maastricht.*

(20') 'Extension to the east is the condition for the stability'

(21') 'The Danish minister of foreign affairs for the newspaper "Lidové noviny"'

(22') 'One of the main topics during the series of discussions of the Danish minister of foreign affairs Nielse Helveg Petersen in Prague was, besides bilateral relations, also the contemporary situation in the European Union.'

(23') 'The Danish road to acceptance of "Maastricht agreement" was not easy at all: the first nationwide voting refused the agreement.'

(24') lit. 'It-was necessary many months of-persuading in-order-to Refl in repeated voting Danes in-the-end by-slight majority expressed for Maastricht.'

('It was necessary to persuade Ø many months, so as, in the end, Danes during the repeated voting by a slight majority agreed with "Maastricht".')

As we can see in above examples, deverbative nouns (e.g. *rozšíření* 'extension', *přesvědčování* 'persuading', *hlasování* 'voting') are very often used without their participants, while author assumes the reader will understand connection between these deletions and their referent from the context. In the same vein as with verbs, we can differentiate the described types of deletions (cf. section 5.1 above) also with deverbative nouns. Considering specific properties of valency frames of nouns, deletions of complementations of nouns should be specified in comparison with those of verbs. We will inquire into these rather subtle differences analysing all participants together, taking into account again coreferential relations between the deleted node and its referent.

5.2.2.1. Transition from textual coreference to coreferential relations without a concrete antecedent

Considering specific and limited properties of surface expression of valency slots of deverbative nouns, the first criterion important for distinction of particular types of deletions with verbs – the typical surface realization of a construction with the respective type of deletion – yields the same results with all participants. The only typical realization of deletions with nouns is the zero position.

For the distinction between deletions with the newly introduced lemma *Unsp* (cf. example (24)) and the cases of General Participant (cf. examples (25) and (26)) the possibility of delimitation of the referent is still very important.

It is typical for sentences illustrating General Participants with nouns that the zero positions refer to all possible Actors, Patients or Addressees; these nouns are used generally, e.g.:

(25) *Tento tuk je vhodný na pečení Gen.*ACT *Gen.*PAT

(25') 'This fat is good for baking'

(26) *Při vaření Gen.*ACT *Gen.*PAT *je třeba zapnout také digestoř*

(26') 'During cooking it is necessary to switch on also the fume chamber'

As was shown in the newspaper article in the section 5.2.2 above, many deverbative nouns are used without their complementations. To understand the connection between the zero position and its referent we must search into the context. We can exactly indicate the referent in case of the actual deletion, as in (20). It is clear from the context that "*rozšíření Evropské unie.*ACT" 'extension of the European Union.ACT' is concerned. So the node with the pronominal lemma should be restored and the attribute COREF gets the lemma "unie".

(20) *Rozšíření na východ je podmínkou stability*

(20') 'Extension to the east is the condition for the stability'

However, with most of the nouns the referent can be only partly delimited from the context, so we propose to mark the restored nodes with the lemma *Unsp*:

(24) *Bylo zapotřebí řady měsíců přesvědčování, aby se v opakovaném hlasování Dánové nakonec těsnou většinou vyjádřili pro Maastricht.*

(24') lit. 'It-was necessary many months of-persuading in-order-to Refl in repeated voting Danes in-the-end by-slight majority expressed for Maastricht.'

('It was necessary to persuade Ø many months, so as, in the end, Danes during the repeated voting by a slight majority agreed with "Maastricht".')

It seems to be obvious that Danes are the referent of the Addressee of the noun *přesvědčování* 'persuading'. We can guess from the context that they were being persuaded that it would be good to vote for the European Union, but we do not know this exactly. Also, we do not know, even not from the context, who was persuading Danes. Danish politicians could be referred by the ACT as persuading them, but again, we do not know it exactly. So, with the noun *přesvědčování* 'persuading' we propose to restore two nodes for unspecified elements, *Unsp*.PAT and *Unsp*.ACT, and one node with the pronominal lemma *on* 'he' and the functor ADDR (the attribute COREF gets the lemma "*Danes*").

In the given sentence one more deverbative noun is used without its complementations, the noun *hlasování* 'voting'. As for the Actor, we can guess from the context, that Danes voted, but, the adjective *celonárodní* 'nationwide' was used as the dependent of the noun *hlasování* 'voting' in the preceding sentence, so we do not know if the author wanted to say that "Danes voted" or "the whole nation voted". Therefore we propose to restore the node for Actor with the lemma *Unsp*. Similarly, we can understand from the context that the matter of the voting was the "*přijetí Maastrichtské smlouvy*" 'acceptance of Maastricht agreement', but again, the author could want to say that "Danes voted for the European Union" - we do not know exactly since the pronoun has not been used. So we propose to restore also the node for Patient with the lemma *Unsp*.

Since the process of substantivization (see section 5.2.1.3 above) can be partial, some deverbative nouns may have as their dependents not only their participants and free modifications but also some modifications typical for nouns denoting substance. When the dependent is expressed by an adjective and it is not any of the free modifications, it is not clear if the function of the dependent is a valency slot of the given noun or just the restrictive attribute. Therefore we propose to restore deleted participants of the given noun and to mark one of them (the one with possible connection to the attribute expressed by

the adjective) with the lemma *Unsp* (especially in case when no other participant of the given noun is expressed at the surface layer of the sentence), e.g.:

(27) *soudní rozhodnutí* 'judicial decision' *Unsp*.ACT

(28) *izraelsko-palestinská jednání* 'Israel-Palestine negotiations' *Unsp*.ACT, *Unsp*.ADDR

(29) *členská evidence* 'membership registration' *Unsp*.PAT.

5.2.2.2. Grammatical coreference: Control with nominalizations

Restoration of all obligatory members of valency frames of postverbal nouns in MC has brought also a lot of interesting examples concerning the cases of deletions with the grammatical coreference. In (30) there occurs two postverbal nouns (i.e. *slib* 'promise' and *omezení* 'reducing') the complementations of which have to be restored. While the two deleted participants of the noun *slib* 'promise', (i.e. nodes for the Actor and the Addressee) represent types of deletion described in the preceding section, the relation between the restored node for the Actor of the noun *omezení* 'reducing' and its antecedent (i.e. the Actor of the noun *slib* 'promise') can be classified as a case of the grammatical coreference.

(30) *Každý vystavovatel musel letos podepsat slib omezení hluku.*

(30') 'This year, every exhibitor had to sign the promise of reducing of noise.'

In the present section we concentrate on nodes which are omitted due to the phenomenon usually called grammatical control with regard to their respective anaphoric relations (cf. Panevová, Řezníčková, and Urešová, 2002). In particular, we extend the notion of control to nominalization and demonstrate how this relation is captured in PDT.

In FGD, on the underlying or tectogrammatical level, control is a relation of an obligatory or an optional referential identity between a *controller* (antecedent) and a *controllee* (empty subject of the nonfinite complement (= *controlled item*)). The controller is one of the participants in the valency frame of the governing verb (Actor (ACT), Addressee (ADDR), or Patient (PAT)). The controlled item functions also as a filler of a dependency slot in the valency frame of the governing verb, being labeled as Patient or Actor. The empty subject of the controlled item may have the function of different dependency relations to its head word (the infinitive): Actor, or, with passivization of the controlled item, Addressee or Patient (cf. Koktová, 1992).

5.2.2.2.1. Classification of verbs of control with controlled infinitive

Panevová and Koktová classify the verbs of control according to the type of its valency frame and to the functions of the controlled infinitive and the controller in the valency frame of the verb of control (see Panevová 1986, 1996, and Koktová, 1992). According to this classification the following basic groups of verbs of control should be recognized (we leave out here some groups with rare types of verbs of control, e.g. verbs with the so-called Slavonic Accusative with Infinitive, e.g. *Viděl Karla přicházet* (lit. 'He saw Charles to-come')):

1. The controlled infinitive functions as Patient: three groups of verbs of control in Czech can be distinguished, namely verbs in the valency frame of which the controller is:

i) ACT (e.g. *Jan se bojí zůstat doma sám* 'John is afraid to stay at home alone')

ii) ADDR (e.g. *Redaktor doporučil autorovi provést několik změn v textu* 'An editor recommended the author to make several changes in the text')

iii) ACT or ADDR (the verb *slíbit* 'promise' with the controller functioning as ACT: e.g. *Jan slíbil matce vrátit se domů před půlnocí* 'John promised his mother to return at home before midnight'; the same verb with the controller functioning as ADDR e.g. *Rodiče slíbili dětem užít si prázdniny ve stanu u rybníka* (lit. 'The parents promised (their) children to enjoy the holidays in a tent by a lake')

2. The controlled infinitive functions as Actor: especially the "predicate" of control (expressed by a copula with an evaluative or modal adjective) is taken into account (e.g. *Je snadné číst tu knihu* 'It is easy to read the book')

3. The controlled infinitive can have also another function, as with raising (e.g. *Viktor se zdá být chytrý* 'Viktor seems to be clever') and the function of attribute (e.g. *Viktor nesmí propást šanci vyhrát* 'Viktor may not miss the occasion to win').

5.2.2.2.2. Extension of verbs of control with "analytical predicates"

The most typical verbs of control (belonging to the group (1)(i)) are the "phase verbs" (e.g. *začít* 'begin', *zůstat* 'stay', *přestat* 'stop'). While describing the phenomenon of control, it seems to be necessary to take into account also more or less synonymous verbs cooccurring with pure modal verbs (the latter are analysed as values of a grammateme in the TGTSs). This position is occupied not only by "modal verbs in the wider sense" (*umět* 'be able', *dovést* 'know how to do sth', *dokázat* 'manage', *zdráhat se* 'hesitate', *odmítat* 'refuse' etc.) but also by "analytical predicates" with a modal meaning (the verb *mít* 'have' plus a certain noun, e.g. *mít schopnost* (lit. 'have an ability'), *dar* (lit. 'have a gift / talent'), *potřebu* 'have an urge to do sth', *příležitost* 'have an opportunity', *šanci* 'have a chance'; the verb *být* 'be' plus a modal adjective, e.g. *být schopen* 'be able', *ochoten* 'be willing', *povinen* 'be obliged').

Also some verbs from other semantic groups of verbs of control can be expressed by an analytical predicate. For example verbs expressing intent, e.g. *hodlat* 'intend', *snažit se* 'try', can be paraphrased by predicates *mít v úmyslu (úmysl), záměr* (lit. 'have an intention'), *mít v plánu (plán)* (lit. 'have a plan'), *mít tendenci* (lit. 'have a tendency') etc.; *být připraven* 'be ready', *odhodlán* 'be determined' etc. (they belong also to the group (1)(i)). Verbs expressing the meaning "*umožnit někomu udělat něco*" 'make it possible for somebody to do something' can be paraphrased by analytical predicates *dát někomu šanci (příležitost) udělat něco* (lit. 'give somebody a chance (an opportunity) to do sth') (these verbs belong to the group (1)(ii)).

5.2.2.2.3. Types of nominalized constructions of control

Panevová (1996) deals not only with controlled infinitive verb structures but also with certain types of nominalizations where the omission of an argument is also based on the control properties of the head (governing) word and must be interpreted as coreferentiality. The group of verbs that offer the possibility for controlled nominalization includes for example verbs such as *přisoudit* 'adjudge', *osočit* 'accuse', *podezírat* 'suspect': *Paní podezírá komornou z krádeže stříbrných příborů* 'The lady suspects the chamber-maid of the theft of silver covers'.

Considering the possibility of a nominalization of both the governing as well as the dependent verb, we deal with four types of constructions of control:

1. The infinitive depends on a finite verb (e.g. *radil nechodit* 'he advised not to go', *slíbil napsat* 'he promised to write');

2. The infinitive depends on a nominalization of a finite verb (e.g. *rada nechodit* 'an advice not to come', *slib napsat* 'a promise to write');

3. The nominalization of the embedded verb depends on a finite verb (e.g. *obvinil někoho z vyvolání problému* 'he charged a person with a raising of a problem', *vyžadoval odpuštění daní* 'he claimed exemption of the taxes');

4. The nominalization of the embedded verb depends on a nominalization of a finite verb (e.g. *obvinění z vyvolání problému* 'an accusation of a raising of a problem', *snaha o podplacení* 'an attempt for corruption').

However, it is necessary to say that not all groups of verbs of control mentioned in section 5.2.2.2.1 allow for their nominalization or for a nominalization of the controlled infinitive:

Verbs of control from the groups (1)(i), (ii), (iii) and (2) may occur in all four types of constructions of control (e.g. verbs *slíbit* 'promise', *vyžadovat* 'require, claim', *snažit se* 'try': *slíbit napsat* 'to promise to write', *slib napsat* 'a promise to write', *slíbit napsání* 'to promise writing', *slib napsání* 'a promise of writing').

Verbs from the group (3) do not allow nominalization in constructions of control.

Verbs such as *přisoudit* 'adjudge', *osočit* 'accuse', *podezírat* 'suspect' may occur only in construction types (3) and (4) (e.g. *podezírat z krádeže* 'to suspect of theft', *podezření z krádeže* 'a suspicion of theft', but **podezírat krást* 'to suspect to steal', **podezření krást* 'a suspicion to steal').

Some of the nouns derived from analytical predicates with the meaning of intent do not always express obligatory grammatical coreference, e.g. *nápad vydat knihu* 'an idea to publish a book' (cf. also Panevová, 1996).

5.2.2.2.4. Coreferential relations in nominalized constructions of control

Nominalized constructions retain those coreferential relations between the controller and the controllee which were realized in constructions with the corresponding verbs of control. Thus, e.g. the nominalized constructions of verbs from the group (1)(iii) mentioned in section 5.2.2.2.1 offer the possibility for the controller to be an Actor or an Addressee. These features are illustrated in the following examples:

1. Constructions in which the Actor of the governing postverbal noun and the Actor of the dependent noun (derived from the predicate expressed by a copula with an adjective) are identical:

(31) *jeho slib poslušnosti* 'his promise of obedience'

(derived from the construction *slíbil, že bude poslušný* 'he promised to be obedient')

The controllee in the valency frame of the dependent noun (i.e. *poslušnost* 'obedience') gets the lemma *Cor* and the functor ACT. Its attributes for coreferential relations get the following values: COREF: *on* 'he', ANTEC: ACT.

2. Constructions in which the Actor of the dependent noun (derived from the predicate expressed by a copula with an adjective) is identical to the ADDR of the governing postverbal noun:

(32) *slib beztrestnosti* 'a promise of impunity'

(derived from the construction *slíbili mu, že bude beztrestný* 'they promised him to be exempt from punishment')

The controllee in the valency frame of the dependent noun (i.e. *beztrestnost* 'impunity') gets the lemma *Cor* and the functor ACT. Its attributes for coreferential relations are filled in by the following values: COREF: *on* 'he', ANTEC: ADDR.

6. Conclusion

The decision on the boundary lines between the different types of deletions linked to the choice among the lemmas of the restored nodes is not an easy task. Our strategy is to mark the difficult cases in a way that allows for their relatively simple identification and thus for preparing resources for further linguistic research. We believe that this solution offers a possibility of further linguistic inquiries into the issues of coreferential relations because it leaves a trace specifying the problematic cases.

Acknowledgment

Research for this paper was supported by the grant of the Czech Ministry of Education LN00A063.

References

Daneš, F. (1971). Větné členy obligatorní, potenciální a fakultativní (Obligatory, Potential and Optional Constituents of the Sentence). In *Miscellanea Linguistica. Acta Universitas Palackiana Olomucensis*. Ostrava, pp. 131-138.

Hajič, J., Hajičová, E., Pajas, P., Panevová, J., Sgall, P., Vidová-Hladká, B. (2001). Prague Dependency Treebank 1.0 (final production label). CDROM CAT: LDC2001T10., ISBN 1-58563-212-0.

Hajičová, E. (1993). Issues of Sentence Structure and Discourse Patterns. *Theoretical and Computational Linguistics*, vol. 2. Charles University, Prague.

Hajičová, E., Ceplová, M. (2000). Deletions and Their Reconstruction in Tectogrammatical Syntactic Tagging of Very Large Corpora. In *Proceedings of COLING'2000*, pp. 278-284, Saarbruecken, Germany.

Hajičová, E., Hajič, J., Hladká, B., Holub, M., P. Pajas, Řezníčková, V., Sgall, P. (2001). The Current Status of the Prague Dependency Treebank. In *Proceedings of the Second Workshop on Text, Speech, Dialogue*, pp. 11-20, Železná Ruda, Czech Republic.

Hajičová, E., Panevová, J., Sgall, P. (2000). Coreference in Annotating a Large Corpus. In Proceedings of LREC 2000, vol. 1., pp. 497-500, Athens, Greece.

Hajičová, E., Sgall, P. (2000). Semantico-Syntactic Tagging of Very Large Corpora: the Case of Restoration of Nodes on the Underlying Level. In *Proceedings of LREC 2000*, vol. 1., pp. 95-98, Athens, Greece.

Koktová, E. (1992). On New Constraints on Anaphora and Control. *Theoretical Linguistics* 18, pp. 102-178.

Kuryłowicz, J. (1936). Dérivation lexicale et dérivation syntaxique. In *Bulletin de la Société linguistique de Paris*, 37, pp. 79-92.

Marková, K., Panevová, J. (in press). Ješčo raz po povodu nulevych elementov v strukture predloženija

Panevová, J. (1973). Věty se všeobecným konatelem (Sentences with a General Actor). In *Studia Slavica Pragensia*, Prague: Charles University, pp. 133-144. Translated in: *Contributions to Functional Syntax, Semantics, and Language Comprehension* (ed. by P. Sgall), Prague: Academia and Amsterdam: John Benjamins, 1984, pp. 203-221.

Panevová, J. (1986). The Czech Infinitive in the Function of Objective and the Rules of Coreference. In *Language and Discourse: Test and Protest* (*Festschrift for P. Sgall*, ed. by J. Mey), Amsterdam – Philadelphia: John Benjamins, pp. 123-142.

Panevová, J. (1996). More Remarks on Control. In *Prague Linguistic Circle Papers*, Vol. 2 (eds. E. Hajičová, O. Leška, P. Sgall, Z. Skoumalová). J. Benjamins Publ. House: Amsterdam - Philadelphia, pp. 101-120.

Panevová, J. (1998). Ellipsis and Zero Elements in the Structure of the Sentence. In *Tipologija, grammatika, semantika. K 65-letiju Viktora Samuiloviťa Chrakovskogo*. Sankt-Peterburg: Nauka, pp. 67-76.

Panevová, J. (2000). Poznámky k valenci podstatných jmen (Notes on the valency of nouns). In Hladká, Z., Karlík, P. (Eds.). *Čeština – univerzália a specifika*, Brno: MU, pp. 173-180.

Panevová, J., Řezníčková, V., Urešová, Z. (2002). The Theory of Control Applied to the Prague Dependency Treebank (PDT). In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks*. Università di Venezia, 2002, pp. 175-180.

Sgall, P., Hajičová, E., Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel and Prague: Academia.

Štícha, F. (1987). Komunikativní a jazykové funkce lexikálního nevyjádření objektu děje ve větě (Communicative and Language Functions of a Lexically Unexpressed Object of Action in the Sentence). *Naše řeč* 70, pp. 184-193.