Deletions and Node Reconstructions in a Dependency-Based Multilevel Annotation Scheme

Jan Hajič, Eva Hajičová, Marie Mikulová, Jiří Mírovský,
Jarmila Panevová, and Daniel Zeman

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czech Republic
{hajic,hajicova,mikulova,mirovsky,panevova,
zeman}@ufal.mff.cuni.cz

Abstract. The aim of the present contribution is to put under scrutiny the ways in which the so-called deletions of elements in the surface shape of the sentence are treated in syntactically annotated corpora and to attempt at a categorization of deletions within a multilevel annotation scheme. We explain first (Sect. 1) the motivations of our research into this matter and in Sect. 2 we briefly overview how deletions are treated in some of the advanced annotation schemes for different languages. The core of the paper is Sect. 3, which is devoted to the treatment of deletions and node reconstructions on the two syntactic levels of annotation of the annotation scheme of the Prague Dependency Treebank (PDT). After a short account of PDT relevant for the issue under discussion (Sect. 3.1) and of the treatment of deletions at the level of surface structure of sentences (Sect. 3.2), we concentrate on selected types of reconstructions of the deleted items on the underlying (tectogrammatical) level of PDT (Sect. 3.3). In Section 3.4 we present some statistical data that offer a stimulating and encouraging ground for further investigations, both for linguistic theory and annotation practice. The results and the advantages of the approach applied and further perspectives are summarized in Sect. 4.

1 Motivation and Specification of Deletions (Ellipsis)

Deletion (ellipsis) in language is a long-standing hard problem for all types of theories of formal description of language, and, consequently, also for those who design annotation schemes for language corpora. As such, this phenomenon present in all languages deserves a special attention both from the theoretical viewpoint as well as with regard to empirical studies based on large annotated corpora. Our contribution is based on a dependency-based grammatical theory, on a multilevel treatment of language system and is supported by language data present in the Prague Dependency Treebank for Czech (PDT); when relevant, we also comment upon the English data of the deep-structure annotation of the Wall Street Journal.[1]

[1] A theoretically-oriented analysis of ellipsis from the point of view of dependency grammar is presented in Panevová, Mikulová and Hajičová, to be submitted for DepLing 2015.

A. Gelbukh (Ed.): CICLing 2015, Part I, LNCS 9041, pp. 17–31, 2015.
DOI: 10.1007/978-3-319-18111-0_2

Ellipsis is generally defined as an omission of a unit at the surface shape of the sentence that is, however, necessary for the semantic interpretation of the sentence. In other words, ellipsis may be regarded as an empty place in a sentence that has not been occupied by a lexical unit. A similar specification is given by Fillmore (2000) who discusses elements that are represented as "understood but missing" and distinguishes Constructionally Licensed Null Instantiation, Indefinite Null Instantiation, and Definite Null Instantiation, as separate ways of cataloguing the "missing" elements. In a similar vein, Kayne (2005, p.v) speaks about silent elements, that is "elements that despite their lack of phonetic realization seem to have an important role in the syntax of all languages".[2]

With language descriptions working with two syntactic levels, one reflecting the surface shape of the sentence, and one representing the level of deep syntactic structure (linguistic meaning), it is possible to consider an establishment of a new element (node in a tree-like representation of the sentence) in the deep structure tree. From this point of view, two situations may obtain: (i) the newly established ("reconstructed") node on the deep level corresponds to an element that as a matter of fact might have been an element (even if perhaps stylistically awkward) of the surface structure but which has been actually "deleted" (we may call this situation a "textual" deletion/ellipsis), as is the case of *John gave a flower to Mary and [he gave] a book to his son*, or (ii) the grammatical structure of the surface shape of the given sentence does not allow for such an insertion but the semantic interpretation of the sentence requires a node to be present in the deep structure (e.g. the controllee in the constructions with verbs of control, such as *John promised to come* has to be interpreted as *John promised that he=John comes*). This type of ellipsis may be called grammaticalized ellipsis.

2 Treatment of Ellipsis in Some of the Advanced Annotation Schemes for Different Languages

There are not very many studies on ellipsis within the formalism of dependency grammar, and even less frequent are general treatments of this phenomenon in annotation scenarios for corpora. However, as the developers of annotation schemes often have to provide instructions how to deal with such a phenomenon, one can observe some commonalities and differences in schemes for individual languages.

One of the most frequent and complicated types of deletion occurs in coordinated structures in which one element of the structure is missing and for its dependents ("orphan") there is no suitable parent node. Several solutions have been adopted:[3] the "orphans" are "lifted" to a position where their head would have been placed and marked by a special label (similar to the label ExD in the analytical level of PDT, see below Sect.

[2] Such a broad specification of ellipsis allows to include under such a heading also cases of movement or shifting or similar kinds of restructuring, as Chaves (2014) duly notes. However, this is not our concern in the present contribution.

[3] For a detailed analysis of the treatment of coordination in most different dependency treebanks and for the taxonomy of these approaches, see Popel et al. (2013).

3.2). Similar to the PDT treatment is that of the Danish Treebank: the "orphan" is placed where the missing parent node would be and is attached to an existing node and marked by a special label. Thus in the tree for *Skær de rensede løg igennem en gang og derefter i mindre stykker på tværs.* [Cut the cleaned onions through once and then into smaller pieces across.] the node for *derefter* [then] is attached to the conjunction *og* and assigned the label <mod> (i.e. there is no copy of the verb *skær*). In a similar vein, the phrases *i mindre stykker* and *på tværs* are attached to the conjunction and labelled as <avobj> and as <mod>, respectively. Had their head verb been present, they would be labeled avobj and mod (without the angle brackets).

A different solution is proposed in the Universal Stanford Dependency scheme,[4] in which the "orphan" is attached by means of a special dependency function called remnant to the corresponding dependent of the non-deleted governor. Thus, e.g. in a sentence corresponding to English *John visited Mary and George Eva*, the node for *George* would "depend" on *John*, and *Eva* on *Mary* (and both *John* and *Eva* on the verb *visited*, with their corresponding dependency relations, e.g. Subj, and Obj respectively); such a treatment can be understood as an attempt to "copy" the node of the expressed verb, but would lead to serious non-projectivities; its advantage is that the reconstruction including the identification of the type of dependency would be straightforward.

Another possibility is to establish an independent NULL node that represents the deleted second verb; the "orphans" are then attached to this newly established node. As far as we can say, there is no reference to the first verb and also there are no copies of the lemma etc. of this first node. One example of an insertion of empty heads is the insertion of the Phantom node in the SYNTAGRUS corpus for Russian; another example is the Turku Dependency Treebank of Finnish (Haverinen et al. 2010).[5] The same is true about the Hindi Treebank (Husain et al. 2010).

In the dependency treebank of Russian, SYNTAGRUS (Boguslavsky et al. 2009)[6], one sentence token basically corresponds to one node in the dependency tree. There is, however, a noticeable number of exceptions, one of which concerns so-called Phantom nodes for the representation of those cases of deletions of heads that do not correspond to any particular token in the sentence; e.g. *ja kupil rubashku, a on galstuk*

[4] The "remnant" analysis adopted in the Universal Stanford Dependencies is discussed briefly in de Marneffe et al. (2014).

[5] E.g. in *Liikettä ei ole, ei *null* toimintaa.* [There is no movement, no action.] the copula *ole* (the negative verb *ei* is attached similarly to negative particles in other languages) is the root which in turn is the head of the node *null* the type of relation being `conj` (a "Stanford" style of coordination). Attached to this null node is the second negative particle *ei* (as `neg`) and *toimintaa* (as `nsubj`).

[6] SYNTAGRUS currently contains about 40,000 sentences (roughly 520,000 words) belonging to texts from a variety of genres and is steadily growing. It is the only corpus of Russian supplied with comprehensive morphological and syntactic annotation. The latter is presented in the form of a full dependency tree provided for every sentence; nodes represent words annotated with parts of speech and morphological features, while arcs are labeled with syntactic dependency types. There are over 65 distinct dependency labels in the treebank, half of which are taken from Meaning-Text Theory (see e.g. Mel'čuk, 1988).

[I bought a shirt and he a tie], which is expanded into *ja kupil rubashku, a on kupil.*PHANTOM *galstuk*. A Phantom node gets a morphological tag by which it is characterized. In the version of SYNTAGRUS discussed in Nivre et al. (2008), out of the 32000 sentences 478 sentences (1.5%) contained Phantom nodes, and there were 631 Phantom nodes in total. Phantom nodes may be introduced also for cases other than coordination: e.g. the missing copula in *Kak #Phantom vasha familija?* [What PHANTOM your name], *bojatsja otvetsvennosti kak cherty #Phantom ladana* [They fear responsibility as devils PHANTOM incense].

A different situation occurs when a sentence element present in the surface is understood as a modification of more than a single element (a shared modification), as in *John bought and ate an apple*. Here *John* modifies the two conjuncts as their subject. Several strategies are applied in different treebanks: in the "Prague" style treebanks the shared modification is attached to the head of the coordination, mostly a node representing the conjunction, and it is marked in some way to be distinguished from other nodes of the coordination; in the "Stanford" and "Mel'čuk" styles the first conjunct of the coordination is considered to be the head of the coordination.[7]

In the German TIGER Treebank[8], the elided (i.e. borrowed, copied) constituents in coordinate clauses are represented by so-called secondary edges, also labelled with a grammatical function. This feature facilitates well-targeted extraction of syntactic trees that embody various types of coordinate ellipsis. (Secondary edges are represented by curved arrows in TIGER tree diagrams.) According to Brants et al. (2004: p. 599), "secondary edges are only employed for the annotation of coordinated sentences and verb phrases". Nevertheless, secondary edges occasionally turn up as parts of non-clausal coordination types; however, ellipsis in non-clausal coordinate structures was not annotated systematically.

Deletions occurring in the so-called pro-drop languages and conditioned by the fact that the occurrence of subjects in sentences can be omitted are treated either by reflecting the surface structure, with no additional node inserted in the representation of the sentence (this treatment is present in the treebanks of Italian, Portuguese and Hindi, and also in the analytical level of PDT and in other "Prague" style treebanks), or a new node is established (depending on the verb that lacks a subject in the surface shape of the sentence) as the subject of that verb and marked by a morphological tag for pronouns. See the Spanish *La mujer toma riendas que _ nunca usó* [The woman

[7] We refer to these two "styles" without describing them in detail but we assume that it is clear from the context which treebanks are referred to.

[8] The TIGER Treebank (Release 2) contains 50,474 German syntactically annotated sentences from a German newspaper corpus (Brants et al., 2004). The annotation scheme uses many clause-level grammatical functions (subject, direct and indirect object, complement, modifier, etc.) represented as edge labels in the sentence diagrams). As reported in Harbush and Kempen (2007), the total of 7,194 corpus sentences (about 14 percent) include at least one clausal coordination, and in more than half of these (4,046) one or more constituents have been elided and need to be borrowed from the other conjunct.

takes reins that [missing:she] never used] (Taulé et al. 2008).[9] A similar approach is reflected in the tectogrammatical level of PDT (see Sect. 3.1 below).

A special category that may be also placed under the notion of ellipsis is represented by independent sentences without a predicate, headings etc. In most treebanks, there is just one non-dependent node, usually labeled ROOT. The label does not distinguish whether this node is a deleted verb, a noun or some other POS. Some treebanks, e.g. the Floresta sintá(c)tica treebank of Portuguese (Afonso et al. 2002), introduces the label UTT for the root of non-verbal sentences. If the root may have more than a single child that would be attached to the missing verb (see the Czech *Majitelé rodinných domků* [omitted:*zaplatí*] *ještě více, pokud topí např. koksem* [The owners of family houses [omitted: will pay] still more if [omitted:they] heat e.g. with coke], no unified treatment can be found.

As can be seen from the above very cursory overview, most annotation schemes that work with a single level of syntactic annotation are inclined to adopt the strategy not to reconstruct nodes in the trees unless such a strategy prevents to capture rather complex sentence structures, or, taken from the opposite angle, they allow for reconstructions of nodes when this reconstruction is evident and well definable (as with omitted subjects and so on). It is no wonder then that in those types of ellipsis in which there is no evident position in the surface structure where a reconstructed node would be placed the treebanks capturing the surface shape of sentences ignore the fact that reconstructions would lead to a more transparent (semantic) interpretation of the sentence. This is the case e.g. with structures with control verbs (e.g. *John decided to leave = John decided that [John=he] leaves*), structures with some type of general modification (e.g. *This book is easy to read*), etc. The usability of a multilevel annotation scheme, with annotations of the surface shape of the sentence and with those of its deep syntactic structure, can be well demonstrated on the parallel annotation of the Prague Czech-English Dependency Treebank (PCEDT),[10] with a two-level annotation of Czech and English; the original English texts are taken from the Penn Treebank, translated to Czech and analyzed, both for Czech and for English, by using the Prague PDT-style of annotation. The same philosophy of annotation has been successfully applied to both sides, namely to reconstruct all missing nodes in the deep syntactic (tectogrammatical) level of annotation that are necessary for a correct interpretation of the meaning of the sentence (see Sect. 3.3 below), except for some very specific types of English constructions that are not present in Czech.

[9] In constructions with modal verbs plus infinitive only a single subject is reconstructed hanging on the infinitive which is also supposed to be the head of the finite verb. Ex. *Puedo afirmar mucho de su trayectoria intelectual* [I can confirm much of his intellectual trajectory].

[10] See Hajič et al. (2012).

3 Ellipsis and Node Reconstruction in the Prague Dependency Treebank

3.1 The Prague Dependency Treebank

The Prague Dependency Treebank (referred to as PDT in the sequel) is an effort inspired by the Penn Treebank; the work started as soon as in the mid-nineties and the overall scheme was published already in 1998 (see e.g. Hajič 1998). The basic idea was to build a corpus annotated not only with respect to the part-of-speech tags and some kind of (surface) sentence structure but capturing also the syntactico-semantic, underlying structure of sentences. Emphasis was put on several specific features:

(i) the annotation scheme is based on a solid, well-developed theory of an integrated language description, formulated in the 1960s and known as Functional Generative Description,

(ii) the annotation scheme is "natively" dependency-based, and the annotation is manual,

(iii) the "deep" syntactic dependency structure (with several semantically-oriented features, called "tectogrammatical" level of annotation) has been conceptually and physically separated from the surface dependency structure and its annotation, with full alignment between the elements (tree nodes) of both annotation levels being kept,

(iv) the basic features of the information structure of the sentence (its topic-focus articulation, TFA) have been included, as a component part of the tectogrammatical annotation level,

(v) from the very beginning, both the annotation process and its results have been envisaged, among other possible applications, as a good test of the underlying linguistic theory.

The Prague Dependency Treebank consists of continuous Czech texts mostly of the journalistic style (taken from the Czech National Corpus) analyzed on three levels of annotation (morphological, surface syntactic shape and deep syntactic structure). At present, the total number of documents annotated on all the three levels is 3,168, amounting to 49,442 sentences and 833,357 (occurrences of) nodes. The PDT version 1.0 (with the annotation of only morphology and the surface dependencies) is available from the Linguistic Data Consortium, as is the PDT version 2.0 (with the annotation of the tectogrammatical level added). Other additions (such as discourse annotation) appeared in PDT 2.5 and in PDT 3.0, which are both available from the LINDAT/CLARIN[11] repository (Bejček et al. 2013).

The annotation scheme has a multilevel architecture: (a) morphological level: all elements (tokens) of the sentence get a lemma and a (disambiguated) morphological tag, (b) analytical level: a dependency tree capturing surface syntactic relations such as subject, object, adverbial: all edges of the dependency tree are labeled with a (structural) tag, and (c) tectogrammatical level capturing the deep syntactic relations: the dependency structure of a sentence is a tree consisting of nodes only for

[11] http://lindat.cz

autonomous meaningful units, called "autosemantic" units or elements; function words such as prepositions, conjunctions, auxiliary verbs etc. are not included as separate nodes in the structure, their contribution to the meaning of the sentence is captured by complex symbols of the autonomous units. The edges of the tree are interpreted as deep syntactic relations such as Actor, Patient, Addressee, different kinds of circumstantial relations etc.; each node carries also one of the values of contextual boundness on the basis of which the topic and the focus of the sentence can be determined. Pronominal coreference is also annotated.[12]

In addition to the above-mentioned three annotation levels in the PDT there is also one non-annotation level, representing the "raw-text". On this level, called word level, the text is segmented into documents and paragraphs and individual tokens are recognized and associated with unique identifiers (for easy and unique reference from the higher annotation levels).

Crucial for the discussion of the issue of ellipsis is the difference between the two syntactic levels, the analytical (with analytical tree structures, ATSs in the sequel) and the tectogrammatical (with tectogrammatical tree structures as representations of sentences, TGTSs) one. In the ATSs all and only those nodes occur that have a lexical realization in the surface shape of the sentence (be they auxiliaries or autonomous lexical units) and also nodes that represent the punctuation marks of all kinds. No insertions of other nodes are possible (with the exception of the root node identifying the tree in the set). In contrast, the TGTS contains nodes for autosemantic lexical units only, but they might be complemented by newly established (reconstructed) nodes for elements that correspond to deletions in the surface structure. A comparison of the ATS and the TGTS of a particular sentence and of TGTS's of most different sentence structures with different types of newly established nodes makes it possible to categorize the reconstructions and analyze them as for their characteristics and statistics, which is the core of our contribution.

3.2 Deletions in the Representation of the Surface Shape of the Sentence

With the approach to ellipsis described above, one issue has to be raised with respect to the ATS. The problem arises if a node representing some element occurring in the surface shape of the sentence has not an appropriate governor in that structure, i.e. there is no node on which the given node depends. A specific label ExD (Extra-Dependency) is introduced to mark such a "pseudo-depending" node. The position of the node with the label ExD in the ATS is governed by specific instructions; basically, it is placed in a position in which the missing governor would be placed (see Sect. 2 for similar approaches).

[12] In the process of the further development of the PDT, additional information is being added to the original one in the follow-up versions of PDT, such as the annotation of basic relations of textual coreference and of discourse relations, multiword expressions etc.

3.3 Reconstructions of Nodes on the Tectogrammatical Level

3.3.1 Treatment of ellipsis on the analytical and tectogrammatical levels of PDT is illustrated in Fig. 1. The ATS structure is displayed in Fig. 1a (left side), where the deletions are not reconstructed (for the pseudo-dependency within a shortened comparative construction the node ExD is used), whereas in the corresponding TGTS in Fig. 1b (right side) the generalized ACT (#Gen.ACT) is established as dependent on the main verb.[13] Another node for #Gen.ACT is newly established in the reduced comparative construction; the full (expanded) shape of the embedded sentence includes both the comparison (CPR) for the whole comparison construction as well as its local modification. Figures 1a, 1b may also help to compare the number of nodes in the ATS structure (with the function words represented by specific nodes) and the number of nodes in TGTS (with the omission of the function words and with the addition of the nodes for the elements deleted on the surface).

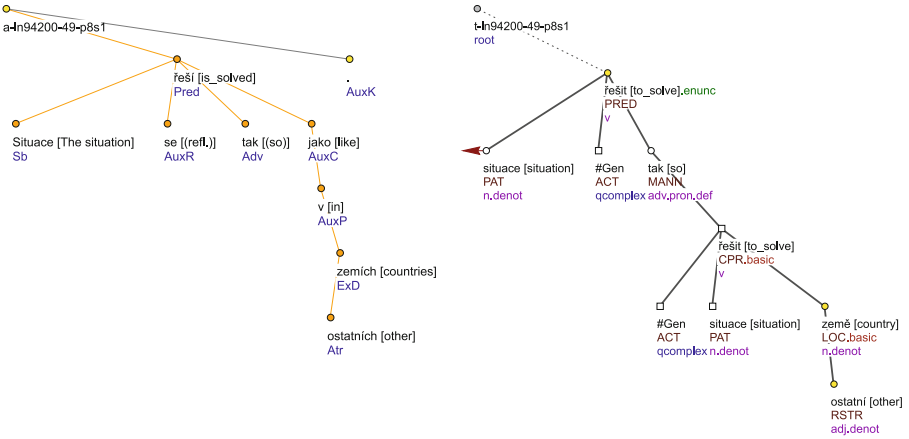

Fig. 1. Situace se řeší tak jako v ostatních zemích.
[The situation is solved like in other countries.]

3.3.2 All syntactically annotated corpora share the problem of reflecting the gaps in coordination constructions. This problem is multiplied by the fact that there exist several types of deletions. In Fig. 2, the omitted noun for one of the conjuncts within coordination in the nominal group is restored by copying the node *podnikání* [enterprise]. (For some properties of the copied nodes, see Sect. 3.4 below.)

[13] The reconstructed nodes in the trees are represented as squares rather than as circles.

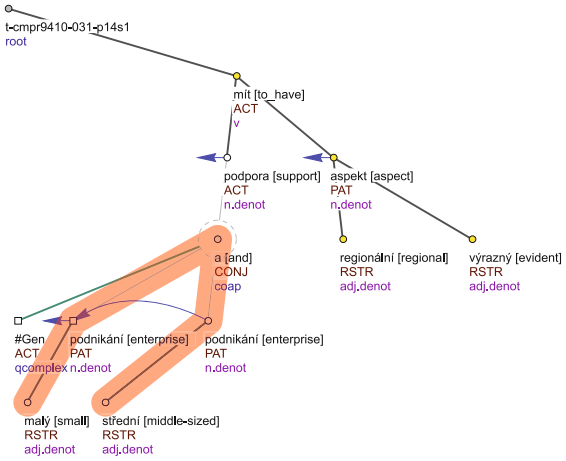

Fig. 2. Podpora malého a středního podnikání má výrazný regionální aspekt.
[Support of small and middle-sized enterprises has an evident regional aspect.]

3.3.3 Fig. 3 exemplifies the PDT treatment of the deletion of the identical predicate in the disjunctive coordination by means of the copied node *dít se* [to_happen]. The shared Actor (Subject) for both clauses is demonstrated here, too. (The treatment of sentence negation present in Fig. 3 is explained below in Sect. 3.3.6.)

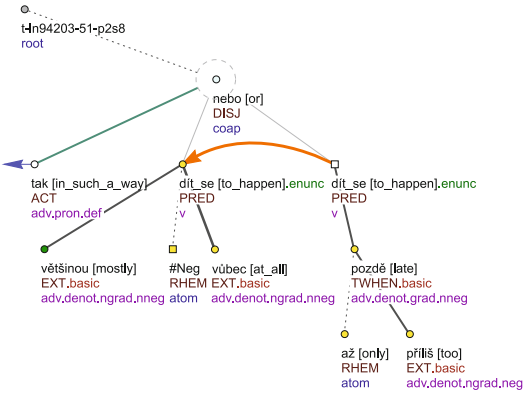

Fig. 3. Většinou se tak neděje vůbec nebo až příliš pozdě.
[Mostly it does not happen in such a way at all or only too late.]

3.3.4 In Fig. 4 the structure of the sentence with missing predicate as the root of the sentence is illustrated. Since the lemma cannot be identified from the context, the node #EmpVerb is established rather than a node with a concrete lemma.

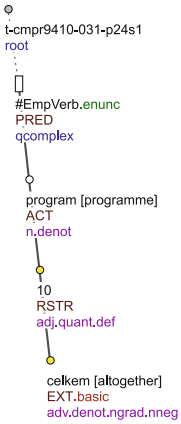

Fig. 4. Celkem 10 programů
[Altogether 10 programmes]

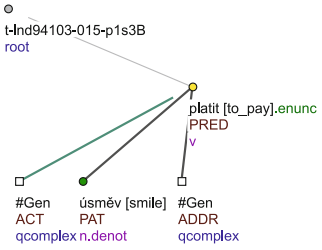

Fig. 5. Za úsměv se platí.
[For smile one pays.]

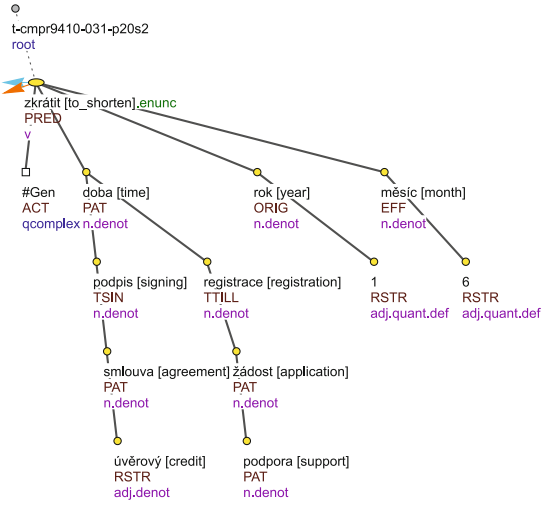

Fig. 6. Byla zkrácena doba od podpisu úvěrové smlouvy k registraci žádosti o podporu z 1 roku na 6 měsíců.
[The time from signing the credit agreement to the registration of the application for support was shortened from 1 year to 6 months.]

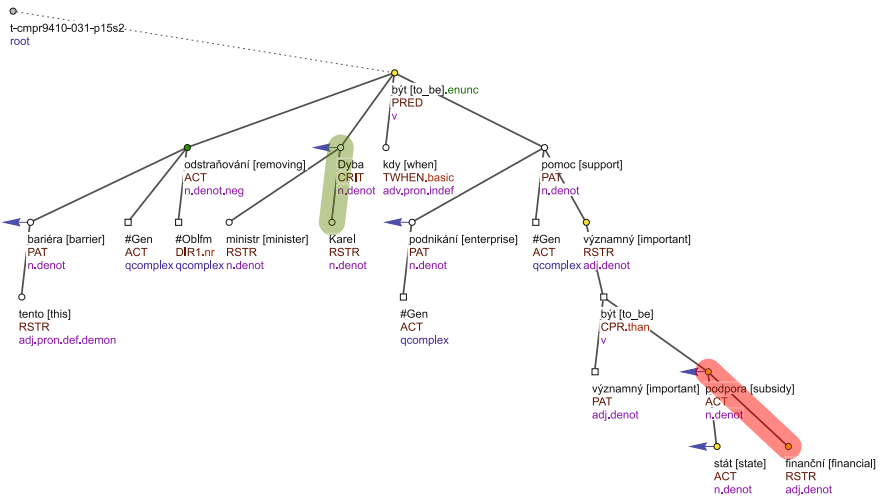

Fig. 7. Odstraňování těchto bariér může být podle ministra Karla Dyby někdy významnější pomocí podnikání než finanční podpora státu.
[Removing of these barriers may be according to minister Karel Dyba sometimes more important support of enterprises than a financial subsidy from the state.]

3.3.5 The generalization of the Actor and of other valency members (participants and some adjuncts) belongs to frequent phenomena in the PDT. For the generalization of ACT (#Gen.ACT), there is a special form in Czech (called deagentization, or, in older tradition, reflexive passive), see Fig. 5 and Fig. 1. General ACT often occurs in passive sentences (see Fig. 6). The generalization of other participants and modifiers missing in the surface shape of the sentence is handled in the TGTS's by added nodes with the lemmas #Gen and their corresponding functions (PAT, ADDR etc.); #Oblfm is used as the lemma for a generalized adjunct. General Actor depending on a deverbal noun is present in Fig. 7, the local modification (LOC) specified as "from where (DIR1)" is annotated here as an obligatory modifier of the noun *odstraňování* [removal].

3.3.6 In Fig. 8, three types of an insertion of a new node are present: (a) The arrow from the newly established node #PersPron.ACT standing for the deleted Actor indicates that the deleted Actor is present in the preceding context. (b) The missing head of the first conjunct *záležitost* [matter] within the noun group is inserted as a copy. (c) In case of sentential negation formed in Czech by a prefix *ne-* attached to the positive form of the verb (*vidí* = he sees, *nevidí* = he does not see) a new node labelled #Neg and attached a functor RHEM is established depending on the verb (the lemma of which is the positive form of the verb). The position of the #Neg node with regard to the verb and other nodes depending on the given verb indicates the (semantic) scope of negation, which in a general case does not necessarily include the verb.

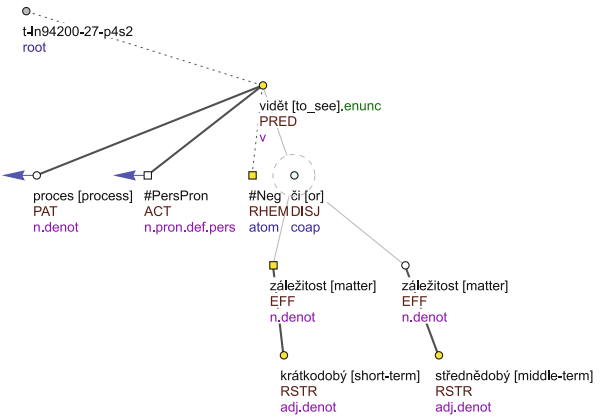

Fig. 8. Proces nevidí jako krátkodobou či střednědobou záležitost.
[He does not see the process as a short-term or a middle-term matter.]

3.3.7 The predicate *lze* [it is possible] is connected with the relation of control. In the given sentence (Fig. 9) the Benefactor functions as the controller (generalized #Gen.BEN). Its Actor fills the role of the controllee and is represented by the node #Cor indicating the grammatical coreference required by the underlying structure of infinitive constructions.

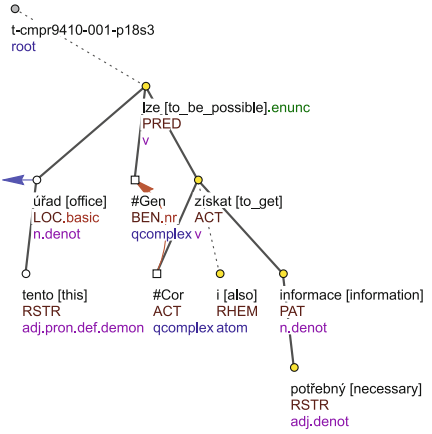

Fig. 9. Na tomto úřadě lze získat i potřebné informace.
[At this office it is possible to get also the necessary information.]

3.4 Some Simple Statistics

The existence of syntactic annotations on two levels of sentence structure allows for some interesting statistical comparisons. Out of the total of 43,955 sentences of the PDT 3.0 training + dtest data (9/10 of the whole PDT) there are 29,243 sentences with

a newly generated node with a t-lemma label of reconstructed nodes and 4,154 sentences with a reconstructed copied node (mostly in coordination structures).

There is a total of 65,593 occurrences of newly generated nodes of the former category (their t-lemma starts with #). The reconstruction of nodes for General Participants prevails rather significantly (see Figs. 5 through 7), followed by cases of reconstructions of nodes mostly for textual deletions in which case the new node labeled as #PersPron has a counterpart in the preceding context (see Fig. 8); these two groups account for 41,136 cases. The next most frequent group (7,476) covers a reconstruction of the controllee in so-called "control" structures (see Fig. 9). The third group relates to negation (7,647 cases), which is more or less a formal reconstruction, though important from the semantic point of view as mentioned above (see Figs. 3 and 8). The categories at the bottom of the frequency list are of a more or less technical character: the label #Forn (1,495) for foreign words or the label #Idph (754) for idiomatic phrases, or they belong to rather specific cases. In between there are three categories that are theoretically biased and given – similarly as #Gen – by the respective verbal valency frames: #Oblfm for semantically obligatory modifications of verbs (1,927 occurrences, see Fig 7), #Unsp for Actor with an unspecified reference without a counterpart on the surface (201 occurrences), and #Rcp for reciprocal constructions (994 occurrences). There is a total of 3,539 nodes for reconstructed root nodes without a lexical label (#EmpVerb, see Fig. 4, and #EmpNoun).

The category of an insertion of so-called copied nodes applies especially in the cases of coordination (see Figs. 2, 3 and 8). In the given set of data, there is a total of 6,799 newly established copied nodes, out of which there are 5,988 cases copied from the same sentence and 811 cases copied from a different sentence. The newly established node is inserted into a position in which it should be placed in the tectogrammatical structure. Both the original node and the copied one refer to the same lexical counterpart on the analytical level (ATS), which is to say that a copied node shares with the "original" node the t-lemma. As for the values of other attributes relevant for the given node, there is a list of those that are copied unchanged together with the t-lemma (e.g. the values for gender, aspect, iterativeness with verbs etc.). However, values other that those given by the list may be changed by the annotator to correspond to their actual value corresponding to the context of the newly established node. This concerns e.g. the values of functors: out of the total of 6,799 newly established nodes 5,027 of them share the value of the functor with the original node, and in 1,772 cases the functors are different. There are 197 pairs of different functors (original – copy), and it is interesting to note that among the first 20 of most frequent pairs (with 1,584 occurrences), in 905 cases (more than 57%) the copied node gets the functor CPR for the relation of comparison (e.g. PRED – CPR, see Fig. 1b).

It was a general principle that any newly established node (i.e. a node not expressed in the surface shape of the sentence) should get the TFA value 't' for a contextually bound element. This default assignment is based on the intuitive assumption that such a node deleted on the surface should refer to a piece of information which has already been in the previous context. However, the annotators were offered the possibility to change the TFA value according to the actual TFA

structure of the sentence. To confirm the validity of the default assignment, we have checked the data in the set of sentences with copied nodes and have found out that in 855 cases the annotators considered necessary to change the default 't' value into the 'f' value (for contextually non-bound). Having checked these sentences carefully, the largest group consisted of coordination of the type *Proces nevidí jako krátkodobou či střednědobou záležitost* [He does not see the process as a short-term or mid-term matter]: here (see Fig. 8 above) the newly established node copies the lemma *záležitost* [matter], shares the functor EFF with the original node (somebody sees something as a matter) and is also a part of the contextually non-bound information (the sentence communicates about the process and says that it is not seen as a short-term and middle-term matter). It follows that the inserted new node *záležitost* [matter] should get the TFA value 'f'. In few cases, the newly inserted node has been considered as a contrastive contextually bound node and marked as such by 'c' see e.g. *I u průmyslové a stavební výroby nejlepší výsledky dociluje polská ekonomika* [Also with the industrial and building production the best results are achieved by Polish economics]. If the reduced coordination constructions are compared with full constructions even on the surface, the element in question would get these values.

4 Summary and Outlook

The problem of ellipsis, the reconstruction of which is triggered by the context or by the type of syntactic structure, is shared by all languages though the rules for the treatment of deletions and their reconstruction may be language specific; this phenomenon represents a difficult issue for syntactic annotation of sentences as well. In our contribution we have focused on the treatment of ellipsis on two levels of syntactic representation based on dependencies, namely on the analytic (surface) one and on the deep (tectogrammatical) one as present in the Prague Dependency Treebank (PDT). We have attempted at a classification of types of ellipsis as reflected in the PDT scenario documenting that each type requires a different treatment in order to achieve an appropriate semantic interpretation of the surface structures in which ellipsis is present. In this way and also by comparing such a scenario with mono-level ones, we wanted to demonstrate the advantages of a corpus scenario reflecting two levels of syntactic structure (surface and deep) separately but with pointers (references/links) which make it possible to search in both levels simultaneously.

The preliminary classification of the types of ellipsis and the data about their frequency drawn from the PDT as presented in this contribution opens new stimuli for more subtle theoretical studies of the relations between surface and deep structure of sentences, of their relations in discourse, and it serves as a great challenge for an explanation of their conditions and sources.

Acknowledgments. We gratefully acknowledge support from the Grant Agency of the Czech Republic (projects n. P406/12/0658, 15-10472S and P406/12/0557) and the Ministry of Education of the Czech Republic (project LM2010013 – LINDAT/CLARIN). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: a treebank for Portuguese. In: Proc. of LREC 2002(2002)
2. Bejček, E., Hajičová, E., Hajič, J., et al.: Prague Dependency Treebank 3.0. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic (2013), http://ufal. mff.cuni.cz/pdt3.0/
3. Boguslavsky, I., et al.: Development of a Russian Tagged Corpus with Lexical and Functional Annotation. In: Proc. of Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovakia, pp. 83–90 (2009)
4. Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic Interpretation of a German Corpus. Research on Language and Computation 2, 597–620 (2004)
5. Chaves Rui, P.: On the Disunity of Right-node Raising Phenomena: Extraposition, Ellipsis and Deletion. Language 90, 834–886 (2014)
6. de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D.: Universal Stanford Dependencies: A cross-linguistic typology. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavík, Iceland, pp. 4585–4592 (2014)
7. Fillmore, C.J.: Silent Anaphora, Corpus, FrameNet and Missing Complements. Paper presented at the TELRI Workshop, Bratislava (November 1999)
8. Hajič, J.: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning, Karolinum, Prague, pp. 106–132 (1998)
9. Hajič, J., Hajičová, E., Panevová, J., et al.: Announcing Prague Czech-English Dependency Treebank 2.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), İstanbul, Turkey, pp. 3153–3160 (2012)
10. Harbusch, K., Kempen, G.: Clausal coordinate ellipsis in German: The TIGER treebank as a source of evidence. In: Proceedings of NODALIDA 2007 (2007)
11. Haverinen, K., Viljanen, T., Laippala, V., Kohonen, S., Ginter, F., Salakoski, T.: Treebanking Finnish. In: Proceedings of TLT9, pp. 79–90 (2010)
12. Husain, S., Mannem, P., Ambati, B., Gadde, P.: The ICON-2010 Tools Contest on Indian Language Dependency Parsing. In: Proc. of ICON 2010, Kharagpur, India (2010)
13. Kayne, R.S.: Movement and Silence, Oxford University Press (2005)
14. Mel'čuk, I.: Dependency Syntax: Theory and Practice. State University of New York Press (1988)
15. Mikulová, M.: Semantic Representation of Ellipsis in the Prague Dependency Treebanks. In: Proceedings of the Twenty-Sixth Conference on Computational Linguistics and Speech Processing ROCLING XXVI, Taipei, Taiwan, pp. 125–138 (2014)
16. Nivre, J., Boguslavsky, I.M., Iomdin, L.L.: Parsing the SynTagRus treebank of Russian. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, pp. 641–648. Association for Computational Linguistics (2008)
17. Panevová, J., Mikulová, M.: Assimetrii mezhdu glubinným i poverxnostnym predstavleniem predlozhenija (na primere dvux tipov obstojatel'stv v cheshskom jazyke). In: Apresjan, J.D., et al. (eds.): Smysly, teksty i drugie zachvatyvajushchie sjuzhety. Sbornik statej v chest'80-letija I. A. Mel'čuk, pp. 486 – 499. Jazyki slavjanskoj kul'tury, Moscow (2012)
18. Popel, M., Mareček, D., Štěpánek, J., Zeman, D., Žabokrtský, Z.: Coordination Structures in Dependency Treebanks. In: Proceedings of ACL, Sofia, Bulgaria (2013)
19. Taulé, M., Martí, M.A., Recasens, M.: AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In: Proc. of LREC 2008 (2008)

http://www.springer.com/978-3-319-18110-3

Computational Linguistics and Intelligent Text Processing
16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I
Gelbukh, A. (Ed.)
2015, XXVIII, 662 p. 109 illus., Softcover
ISBN: 978-3-319-18110-3